applied sciences

MDPI

Article

# An On-Line and Adaptive Method for Detecting Abnormal Events in Videos Using Spatio-Temporal ConvNet

**Samir Bouindour [1,*], Hichem Snoussi [1], Mohamad Mazen Hittawe [2] and Nacef Tazi [1] and Tian Wang [3]**

1 ICD-LM2S, CNRS, University of Technology of Troyes, 10000 Troyes, France; hichem.snoussi@utt.fr (H.S.); nacef.tazi@utt.fr (N.T.)
2 King Abdullah University of Science and Technology, CEMSE, Thuwal 23955-6900, Saudi Arabia; mohamad_mazen.hittawe@utt.fr
3 School of Automation Science and Electrical Engineering, Beihang University, Beijing 100083, China; wangtian@buaa.edu.cn
* Correspondence: samir.bouindour@utt.fr

Received: 24 January 2019; Accepted: 18 February 2019; Published: 21 February 2019


**Abstract:** We address in this paper the problem of abnormal event detection in video-surveillance. In this context, we use only normal events as training samples. We propose to use a modified version of pretrained 3D residual convolutional network to extract spatio-temporal features, and we develop a robust classifier based on the selection of vectors of interest. It is able to learn the normal behavior model and detect potentially dangerous abnormal events. This unsupervised method prevents the marginalization of normal events that occur rarely during the training phase since it minimizes redundancy information, and adapt to the appearance of new normal events that occur during the testing phase. Experimental results on challenging datasets show the superiority of the proposed method compared to the state of the art in both frame-level and pixel-level in anomaly detection task.

**Keywords:** abnormal event detection; deep learning; Convolutional Neural Network; unsupervised learning; online; adaptive systems

## 1. Introduction

Abnormal event detection and localization is a challenging and exciting task in video monitoring. Indeed, the security context in recent years has led to the proliferation of surveillance cameras, which generate large amounts of data. This flow of CCTV images creates a lack of efficiency of human operators. Moreover, studies show that they can miss up to 60% of the target events when they monitor nine or more displays [1]. In addition, after only 20 min of focus, the attention of most human operators decreases to well below acceptable levels [2]. This can lead to potential security breaches, especially when monitoring crowded scene videos.

A possible solution to this problem is the development of automated video surveillance systems that can learn the normal behavior of a scene and detect any deviant event that may pose a security risk.

Abnormal event detection, also known as anomaly detection, can be defined as a spatial temporal recognition problem, taking into account that the event to be recognized is not present in the training phase. In the context of video surveillance systems, anomalies are formed by rare shapes, motions or their combinations. The main challenge in abnormal event detection 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 initial studies in abnormal event detection focused on trajectory analysis [3–5], where a moving object is considered as abnormal if its trajectory doesn't respect the fitted model during the training period. The main limits of such method are the sensitivity to occlusion and the effectiveness of detecting abnormal shapes with normal trajectories. These methods can be used in detecting deviant trajectories in scenes containing few objects but not achieve satisfactory performance for other applications.

Other methods such as low level local visual features [6–8] tried to overcome the limits of trajectory analysis and construct models by using handcrafted feature extractors. Among these methods, low local features such as histogram of oriented gradient (HOG), and histogram of optical flow (HOF) have been used to model the background and to construct the template behavior. However, these methods are usually specific to a given application and are not optimal for complex events. Besides, they don't link between local patterns, since local activity patterns of pixels is not efficient for behavior understanding. More complex methods used the concept of Bag of Video words (BOV) by extracting local video volumes obtained either by dense sampling or by selecting points of interest to construct the template behavior. However, the relationship between video volumes is often not taken into account. Derivatives of these methods [9] attempted to enhance the previous models by using not only the local region, but also the link between these regions for the overall understanding of the events. Usually the complexity of these methods makes them inefficient and time consuming for detection of abnormalities in crowded scenes.

Nowadays, deep learning has aroused the interest of the scientific community and works have been carried out in several fields including agriculture [10], biology [11] and economics [12]. More specifically, deep learning based methods have demonstrated a high capacity on image processing [13,14], which led to the use of supervised methods for the anomaly detection. These methods are generally based on the use of convolutional neural networks (CNNs) [15]. The main drawback of the supervised methods is the use of both normal and abnormal examples in the training phase which makes them not usable in real-world application for video surveillance, because it is very difficult to identify and reproduce all the abnormal events.

Other deep learning based works [16–18] have achieved good performance on anomaly detection datasets. These methods use unsupervised learning and their learning process do not require normal and abnormal training examples, which makes them suitable for abnormal event detection. In this article, we propose an unsupervised on-line adaptive method based on deep learning for the detection and localization of abnormal events. The main contributions of this paper are as follows:

- We adapt a pretrained 3D CNN to extract robust feature maps related to shapes and motions which allow us to detect and localize complex abnormal events in non-crowded and crowded scenes.
- We propose a new method of outliers detection based on the selection of vectors of interest to construct a balanced distribution. This robust classifier is able to represent all normal events (redundant and rare ones) during the training phase, detects abnormalities and adapt to the appearance of new normal events during the testing phase.

The rest of this paper is presented as follows: In Section 2, we present a brief state of the existing 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

In this section, a review of the current main methods is presented.

Initial studies gave attention to the use of the trajectory for anomaly detection [3–5,19–21]. Calderara in [19] represented trajectories as sequences of transitions between nodes in a graph to detect abnormal movements. In another work, Morris [20] combined the trajectory with Gaussian mixture and hidden Markov models to build activity models. Antonakaki [21] proposed an approximation algorithm implementing a hidden Markov model as one-class classifier, which can decide whether a given trajectory is normal compared to a model.

Others works [6–8,22,23] were devoted to the extraction and analysis of low-level local features. Ermis in [6] generated behavior clusters based on behavior profile of a given pixel to construct probabilistic models. Reddy in [7] used optical flow to characterize the object's motion and proposed an adaptive codebook to obtain texture features. In the same process, Wang [8] combined a histogram of optical flow orientation with one class support vector machine (SVM) for detecting abnormal motions. Boiman and Irani [22] presented an approach based on spatio-temporal volumes and considered that if the event reconstruction using only the previous observations is impossible then this event is classified as abnormal. Roshtkhari [9] proposed a method which is the improvement of [22], by using a codebook to group similar spatio-temporal volumes to reduce the dimension of the problem. This proposed method allows to reduce the computational complexity, and becomes applicable to real-time video analysis. Xiao in [23] proposed a sparse semi-non-negative matrix factorization (SSMF) to learn local patterns, and used a histogram of non-negative coefficients (HNC) as local descriptors. Li et al. in [24] used a dynamic texture (DTs) to design models of normalcy over both space and time. Spatial and temporal information is modeled with a mixture of DTs (MDT).

Recently, various methods based on deep learning were applied for anomaly detection [15–17,25–27]. Zhou [15] proposed a method for detecting and locating abnormalities based on spatiotemporal convolutional neural network. In this paper, the authors used a 3D-CNN to extract features related to motions and shapes. Although the results presented were encouraging, the supervised nature of this algorithm did not allow its use in real world systems monitoring. Indeed, both normal and abnormal events were used as training examples, which is not suitable in the field of video surveillance since it is very hard to reproduce all abnormalities. Sabokrou et al. in [26] used three initial layers of pre-trained 2D-CNN combined with a new convolutional layer, whose kernels were learned using sparse auto-encoder for abnormal event detection. One should notice here that the 2D convolution does not sufficiently take into account the temporal information. Same author in [27] used local and global descriptors generated with sparse auto-encoder to capture video properties. Hasan in [16] used a fully convolutional auto-encoder (AE), the learned AE reconstructs normal events with low error and give higher reconstruction error for abnormal events. For the same purpose, Chong [25] proposed an AE with two parts, 2D convolutional layers for spatial features and use convolutional long short term memory for temporal information. Ravanbakhsh in [17] proposed to use a generative adversarial nets (GANs), which are trained using normal frames and their corresponding representation generated by optical flow. In the testing phase, the real data are compared with the reconstruction generated by GANs, the use of optical flow in the detection of abnormalities can be time consuming. Xu et al. in [28] presented an unsupervised deep learning framework for abnormal event detection in complex scenes. The proposed method is based on stacked denoising autoencoders (SDAE) to extract features, and multiples one-class-SVM to predict score of abnormalities. Thus, they used a fusion strategy for the final decision. Ravanbakhsh et al. in [29] combined a fully convolutional network (FCN) with a binary quantization layer that was trained by an iterative quantization (ITQ) for obtaining binary bit maps. Using this technique, the authors could get a set of binary codes for every video volume. At the end, a histogram representing the distribution of binary codes was generated.

Methods for video features extraction were developed recently [30–32] and have achieved promising results in action recognition. The advantage of these methods lies in their ability to learn spatiotemporal information. Despite the fact that these methods are mostly supervised, it can be adapted effectively for abnormal event detection task.

Our combination of pretrained 3D-CNN and the proposed classifier is an improvement of our previous work [18] where we introduced a method for abnormalities detection based on a 2D-CNN combined with a one-class SVM [33,34]. Although the results generated were acceptable, the motion information is reduced by the 2D convolution. Moreover, the setting of an SVM and more specifically the choice of the kernel is complex, it is also admitted that the efficiency of an SVM is strongly impacted by the size of the data [35]. For these reasons, we propose in this paper a 3D-CNN combined with a new classifier to overcome these limits: From one side, the 3D-CNN is more suitable for spatiotemporal

abnormal event detection. From another side, the new classifier is robust to the evolution of the training dataset size, it also takes into account the normal events that occur rarely during the training period and can adaptively learn newly observed normal events in the testing phase. Thus, we obtain a method directly usable for real-world applications. To sum up, Table 1 summarizes the main methods presented in this section, their advantages and drawbacks.

**Table 1.** Comparison of some methods of abnormal events detection in terms of advantages and disadvantages.

| Method | Advantages | Disadvantages |
| --- | --- | --- |
| Trajectory analysis | Effective for detecting deviant trajectories in non-crowded scenes | Occlusion-sensitive especially in crowded scenes. Inability to detect abnormal shapes without deviant trajectories |
| Handcrafted features | Effective for simple local shape or motion | Does not link between local patterns, not optimal for complex events |
| Deep learning: Supervised | Effective for behavior understanding<br>High capacity on image processing | Need to use normal and abnormal examples during the training phase |
| Deep learning: Unsupervised | Effective for behavior understanding<br>High capacity on image processing<br>Does not require normal and abnormal training examples | - - - |

## 3. The Proposed Framework

We present in this section the proposed abnormal event detection method. We start by giving a brief description of the 3D residual convolutional networks and we explain the benefits of using these networks comparatively with 2D plain networks. After this, we give details of the two main stages of our method; feature extraction using modified version of a pre-trained 3D residual convolutional network [31], and the detection of outliers using the proposed classifier. Thus, this method allows us to detect abnormal events, prevents the marginalization of rare normal events and adapt to the appearance of new normal events during the testing phase.

### *3.1. 3D Residual Convolutional Networks*

The 2D-CNNs perform convolution and pooling operations only spatially, while 3D-CNN is used for spatio-temporal representations. Features from 3D convolutional networks encapsulate information related to shapes and motions in sequence of images, which can be useful for abnormal event detection. Besides, Tran et al. in [30] proved that their C3D features are separated remarkably compared to 2D based model, which can be desirable for videos representations. Figure 1 shows how it can be effective to perform clustering with features generated by C3D than those generated by ImageNet [36] on UCF101 dataset.

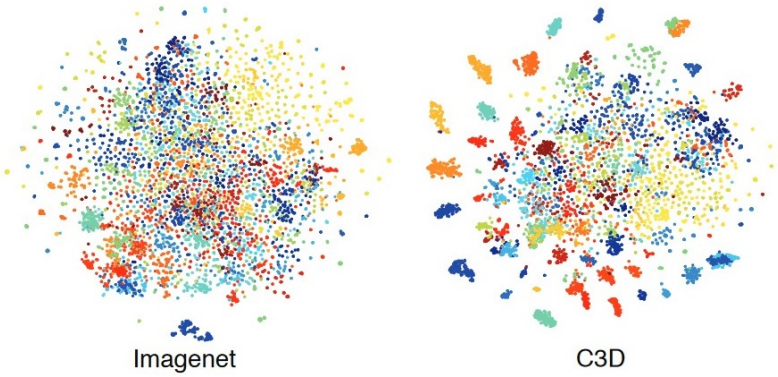


**Figure 1.** Feature embedding visualizations of Imagenet and C3D on UCF101 dateset using t-SNE method [37], this figure was extracted from [30].

The 3D-CNNs proposed in [30–32] proved to be highly competitive compared to different 2D networks in various video analysis tasks, such as scene, object or action recognition.

The use of pretrained 2D convolutional network such as features extractors for abnormal event detection was recently developed by [18,26]. However, it was proven in [30], that image-based feature extraction is not suitable for action recognition, since the 2D convolution does not consider temporal relationship between successive frames. For instance, in [26], a sequence of frames was used as an input of the CNN. However, the convolution kernels in the 2D-CNN does not allow consideration of temporal information. i.e., after the first convolutional layer, the movement information is destroyed by the 2D convolution. That is why the use of multiple frames as an input of the 2D neural networks is not adequate for temporal features extraction.

Furthermore, when training a network with a very large number of layers, the gradient required to update the weight with backpropagation becomes smaller when reaching earlier layers. This is the so-called vanishing gradient problem. As a result, as the network goes deeper, its performance gets saturated or even starts degrading rapidly. That is why He et al. in [14] proposed a novel architecture based on residual blocks (see Figure 2). The original output mapping F(x) is readjusted into $F(x) + x$. It is thus easy to optimize the residual mapping than the first one.

This network named "ResNet" is formed of 152 layers. It has obtained the first place in the Large Scale Visual Recognition Challenge (ILSVRC) 2015 challenge with a classification error of 3.6%.

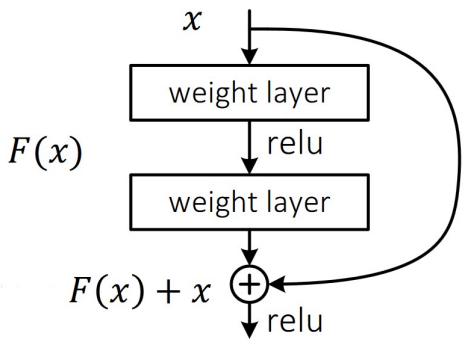


**Figure 2.** Residual learning.

### 3.2. Anomaly Detection

The 3D CNN used in [31] is a residual network pre-trained on Sports-1M dataset containing more than 1.1 million sports videos. This allows it to extract high performed spatio-temporal features from video sequences compared to 2D neural networks used in previous papers [18]. In the following, we present the proposed methodology for abnormal event detection. It is based on two main stages:

Features extraction with modified version of pre-trained Res3D [31] and outliers detection with our classifier.

### 3.2.1. Features Extraction

The proposed methodology consists of using volumes of three consecutive frames. For each frame "$F_t$" we used the volume F: $\{F_t; F_{t-1}; F_{t-2}\}$ as input. As applied in [18], we did not use the complete CNN "from end to end", but we used the four initials 3D residual blocks (16 convolutional layers) as feature extractors. The architecture of this network is presented in Figure 3. Thus, instead of getting an output of one feature vector, the resultant map is a matrix of 841 feature vectors of 256 dimensions. In this matrix, each raw represents a small patch in the input frame. This is explicitly shown in Figure 4.

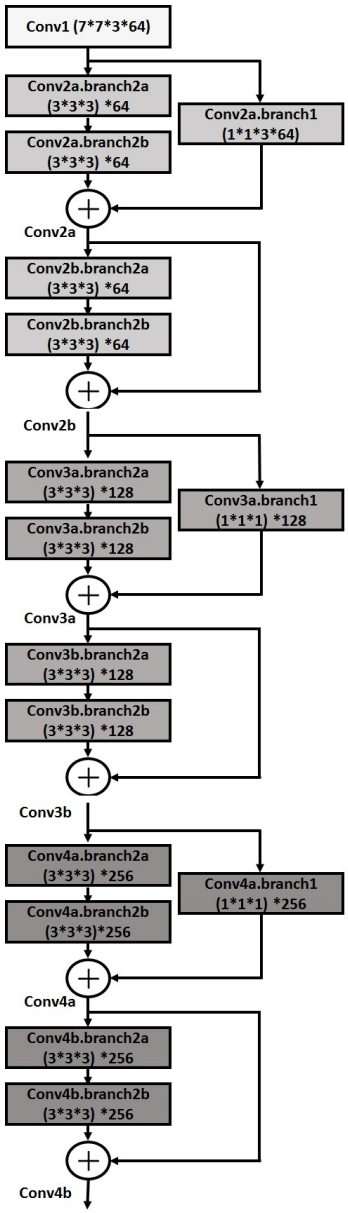


**Figure 3.** The four initial 3D residual blocks of Res3D.

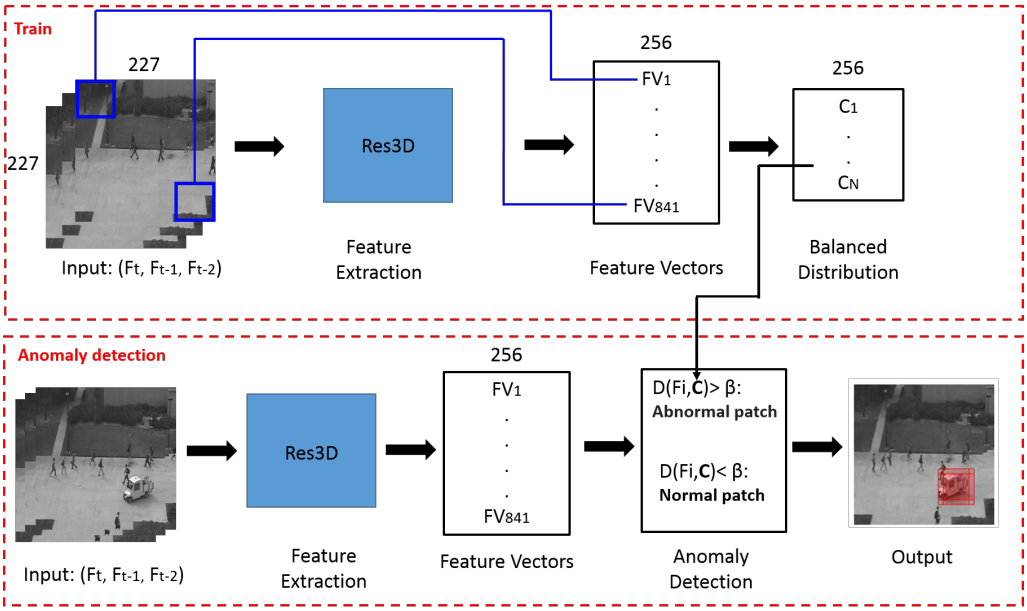


**Figure 4.** Global scheme of the proposed framework.

### 3.2.2. Classifier

CCTV videos are characterized by a redundancy of information in normal scenes. Building the event model by using all vectors obtained for each frame during the training phase would skew the distribution by weighting redundant elements and marginalizing rare ones that would lead to misinterpretations and generate confusion between rare events and abnormalities in the detection process. That is why we proposed in this paper an online selection of vectors of interest. It allows definition of a balanced distribution able to represent all normal events, including rare ones, since the representation of redundant and rare events reaches an equilibrium. Thus, all normal events have the same importance during the detection process so that false alarms would be avoided.

In this way, we were able to construct a robust classifier, able to minimize redundancy, and avoid the marginalization of rare events and reduce false alarms.

Algorithm 1 shows the construction of the balanced distribution. Thus, to select our representative vectors, we declare "N" first vectors from the first frame as vectors of interest. Then, for each new vector "$X_i$" of the processed frame, we calculate a string metric "$dist_i$" based on Mahalanobis distance between "$X(i)$" and the "N" vectors of the distribution. This is represented in the following equation:

$$dist_i = (X(i) - moy) * Q * (X(i) - moy)' \quad (1)$$

where "$moy$" and "$Q$" are the mean and the inverse of the covariance matrix of the distribution, respectively. The Mahalanobis distance considers the correlation between data variables since it is calculated using the inverse of the covariance matrix [38]. Using this distance, the probability that a test point belongs to a set is characterized not by its euclidean distance but by a metric tacking into account the distribution of the data.

In our case, if this distance exceeds a threshold "$\alpha$", the vector is selected to be a vector of interest, then we update "$moy$" and "$Q$". Otherwise, the vector is not considered for the detection of abnormalities. We then continue the same process for each new arrival frame in the training phase. We thus assess the relevance of each vector to be included in the balanced distribution. Finally, this distribution is pruned to eliminate among the first "$N$" elements those that are redundant, which ultimately generates "$M$" different vectors of interest that are considered as the representation of video volumes. This method allows us to compute our algorithm in real time.

Furthermore, this method represents each redundant normal event by a single vector in order to have an equilibrium in point of interest data. It then prevents the marginalization of normal events

that occur rarely during the training process. Thus, every event has the same importance in abnormal detection procedure.

For the testing phase, each patch, represented by a vector obtained by the feature extraction mentioned above is assessed using the same principle. Indeed, we measure the similarity between the vector representing the patch and the balanced distribution. If the measure exceeds a threshold "$\beta$", the vector is considered as an outlier. Thus, the considered patch is labeled as abnormal. The detailed procedure is presented in Algorithm 1. All the thresholds ($\alpha$, $\beta$ and $\eta$) were selected empirically. While "$\alpha$" and "$\eta$" were selected manually according to their effects on the size of the balanced distribution, "$\beta$" was selected in order to obtain the best results (minimum of false alarms and misdetections) in the monitored scenes.

CCTV also faces another challenge due to the constant evolution of the environment that can be characterized as a dynamic background, appearance of new interactions or temporary authorization of abnormal events. These changes can generate false alarms that disturb the interaction with the users of the anomaly detection algorithm.

For instance, a single pathway that is changed to a normal road or the introduction (or removal) of an element in the background. Another example can be the authorization of temporary interactions in an environment such as maintenance works that may create false alarms due to the use of special machines and uncommon tools.

To the best of our knowledge, this problem was skipped in previous algorithms. That is why we proposed an adaptive method robust to any evolution of the surveilled scene.

As seen in Figure 5, this method allows to reconsider false alarms obtained during the test process, i.e., if a patch is declared as a false alarm, the operator can declare its feature vector as a vector of interest to be included in the balanced distribution. This will avoid its future detection as abnormal.

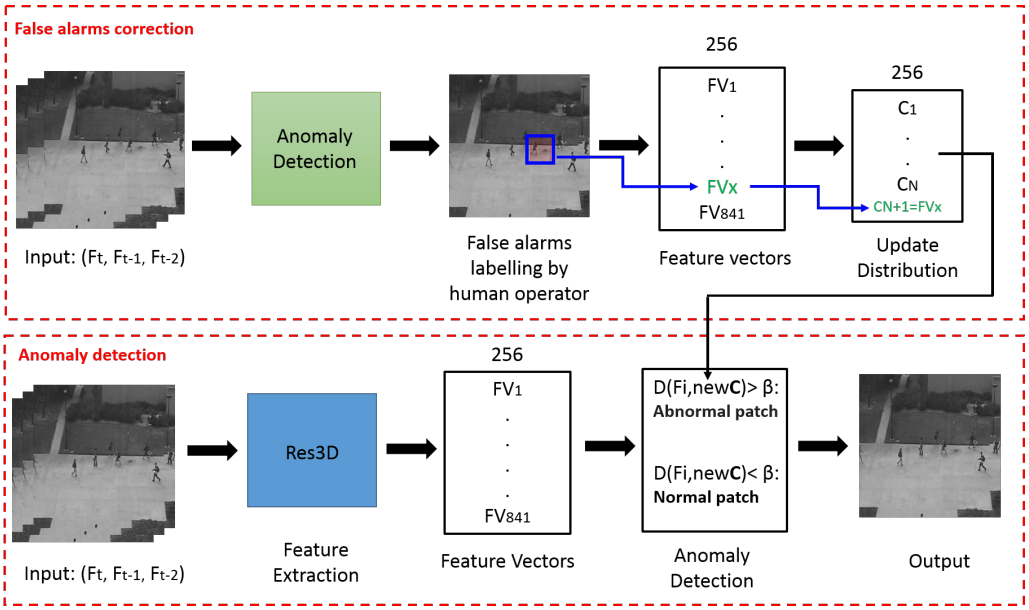


**Figure 5.** Robustness to false alarms.

**Algorithm 1:** Construction of a balanced distribution and abnormal event detection.

```
Selection of Vectors of Interest;
F=[F_i; F_{i-1}; F_{i-2}];
X=FeatureExtraction (F);
C=X(1:N);
moy=mean(C);
Q=inv(cov(C));
for i=N:841 do
    dist_i=(X(i)-moy)*Q*(X(i)-moy)';
    if sqrt(dist_i> α) then
        C(size(C,1)+1)=X(i);
        moy=mean(C);
        Q=inv(cov(C));
    end
end
for each new frame F_i do
    F=[F_i; F_{i-1}F_{i-2}];
    X=FeatureExtraction (F);
    for i=1:841 do
        dist_i=(X(i)-moy)*Q*(X(i)-moy)';
        if sqrt(dist_i> α) then
            C(size(C,1)+1)=X(i);
            moy=mean(C);
            Q=inv(cov(C));
        end
    end
end
Distribution prunning;
moy=mean(C);
Q=inv(cov(C));
for i=1:size(C,1) do
    dist_i=(C(i)-moy)*Q*(C(i)-moy)';
    if sqrt(dist_i< η * α ; (0<η<1)) then
        C(i)=[ ];
    end
end
Abnormal event detection;
for each new frame F_i do
    moy=mean(C);
    Q=inv(cov(C));
    for i=1:841 do
        dist_i=(C(i)-moy)*Q*(C(i)-moy)';
        if sqrt(dist_i> β) then
            Patch_i is Abnormal;
        end
    end
end
```

Furthermore, as any other methods of abnormal event detection, misdetections could also occur while using the proposed framework. That is why we propose as a next step its improvement into a semi-supervised model. Indeed, when the model generates misdetection, the CCTV operator can declare it as abnormal. i.e., the feature vector representing this event will be used to create a second distribution representing misdetected events. Thus, one can get two distributions representing normal and misdetected events. Consequently, to labelize a new event as normal, its feature vector should respect two conditions: The similarity between this new feature vector and the original distribution representing normal events should be established, and we should get a non-similarity with the new distribution representing the misdetected events. Thus, this operation will decrease misdetections and increase the performance of our proposed method. The implementation of this semi-supervised model will be developed in future work.

## 4. Results and Comparisons

In order to implement our methodology, we used Matlab and *caffe* for the deep learning part [36]. We also implement our algorithm in C++ for real time detection.

We tested our methodology on UCSD Ped2 Anomaly Detection Dataset and we apply it also for a real case of laboratory surveillance "CapSec dataset". The local dataset "CapSec" was constructed in our laboratory using a fixed camera. It represents daily behavior of students and researchers inside the laboratory. The image resolution is $1280 \times 720$ pixels. It has training folders that contain only normal events and testing folders which also contains anomalies like falling people and objects that are not present in the training phase. One can see in Figure 6 some qualitative results. UCSD Ped2 "(http://www.svcl.ucsd.edu/projects/anomaly)" represents complex behaviors including pedestrians. It contains 16 and 12 folders for training and testing phases resepctively. The training folders contain only normal events of pedestrians. However, the test folders also contain abnormalities such as skaters, bikes and cars. Abnormal events in this dataset vary in type, size and occurrence. It contains many occlusions and the image resolution is $240 \times 360$, which is low and can complicate the detection process.

While using this dataset, we evaluate two scenarios: "SC1" when we consider a balanced distribution for every folder of the Ped2 dataset, and another scenario "SC2" when we develop a distribution for the whole dataset. The objective of this experiment is to prove the robustness of the proposed method when increasing the number of images during the training phase.

The first results show the effectiveness of the proposed method. The qualitative results are presented in Figure 7 when all abnormalities are detected.

One can see in Table 2 how the introduction of vectors of interest reduces the number of feature vector representing normal events. Consequently, while using the first scenario of the proposed method, we apply a classifier (one classifier per folder) with less than one percent of features vectors selected as vector of interests. In the second scenario, when we consider a unique classifier for all folders containing 1920 training images, we get only 1569 vectors of interest. This number is relatively stable and low even if the training images number increases. This shows how the proposed methodology is robust when facing high number of frames. Consequently, it can be applied for real-world applications of video surveillance.

To assess the performance of the proposed method and provide quantitative results, we used the Equal Error Rate of Frame Level (EERFL), the Equal Error Rate of Pixel Level (EERPL) and the Receiving Operating Characteristic (ROC). A comparison with state of the art methods is provided in Table 3. Our method gets respectively an EERFL and EERPL of 6.25% and 9.82% for the first scenario "SC1", and for the second scenario "SC2" 7.45% and 9.63% respectively, which outperforms all other state of the art methods.

Figure 8 shows the receiver operating characteristic (ROC) for the UCSD Ped2 dataset , plotted as a function of the detection threshold for both scenarios.

**Figure 6.** Detection of abnormal events. (**a**) Falling person; (**b**) Detection of multi-falling people; (**c**) Falling person in presence of walking one; (**d**) Detection of an object assumed as abnormal; (**e**) Falling person in presence of crouched person; (**f**) Falling person.

**Table 2.** Effect of the selection of vectors of interest on element reduction. For each folder of the dataset, one can see the number of frames (NB-Frames), number of feature vectors (NB-FV) and number of vector of interest obtained by our method (NB-VI).

| Folder | NB-Frames | NB-FV | NB-VI |
|---|---|---|---|
| F1 | 120 | 100920 | 934 |
| F2 | 150 | 126150 | 923 |
| F3 | 150 | 126150 | 909 |
| F4 | 180 | 151380 | 1215 |
| F5 | 180 | 151380 | 1127 |
| F6 | 150 | 126150 | 1187 |
| F7 | 150 | 126150 | 1107 |
| F8 | 120 | 100920 | 981 |
| F9 | 180 | 151380 | 1134 |
| F10 | 180 | 151380 | 1177 |
| F11 | 180 | 151380 | 1109 |
| F12 | 180 | 151380 | 1108 |
| SC2 | 1920 | 1614720 | 1569 |

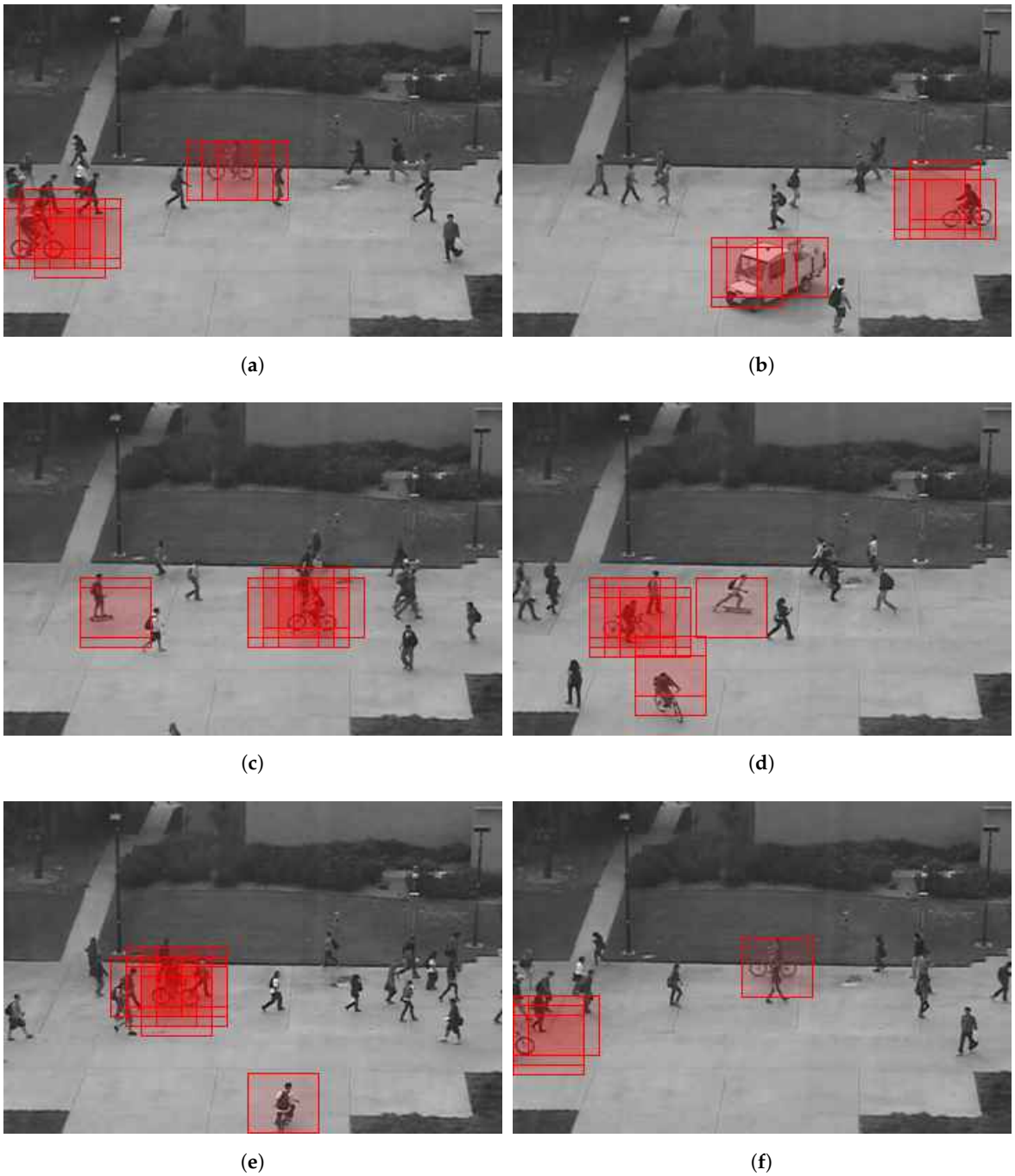

**Figure 7.** Detection of abnormal events in Ped2 Dataset. (**a**) Multiple bicycles detected; (**b**) Detection of multiple targets (bicycle and car); (**c**) Detection of multiple targets (bicycle and skater); (**d**) Detection of multiple targets (bicycles and skater); (**e**) Multiple bicycles detected; (**f**) Detection of multiple targets (bicycle partially obstructed and one wheel).

**Table 3.** Equal Error Rates for frame and pixel level comparisons on Ped2 Dataset. The EER values for the different methods were extracted from the literature.

| Methods | EERFL | EERPL |
|---|---|---|
| Mehran [39] | 42 | 80 |
| Adam [40] | 42 | 76 |
| Bertini [41] | 30 | / |
| Kim(MPCCA) [42] | 30 | 71 |
| Zhou [15] | 24.40 | / |
| Mahadevan(MDT) [43] | 24 | 54 |
| Hasan [16] | 21.7 | / |
| Reddy [7] | 20 | / |
| Sabokrou [27] | 19 | 24 |
| Li [24] | 18.50 | 29.90 |
| Ravanbakhsh [29] | 18 | / |
| Xu (AMDN double fusion) [28] | 17 | / |
| Sabokrou [44] | 15 | / |
| Ravanbakhsh (GAN) [17] | 14 | / |
| Boiman(IBC) [22] | 13 | 26 |
| Roshtkhari(STC) [9] | 13 | 26 |
| Chong [25] | 12 | / |
| Tan Xiao [23] | 10 | 17 |
| Sabokrou [26] | 11 | 15 |
| Sabokrou [45] | 8.2 | 19 |
| Sabokrou [46] | 7.5 | 16 |
| **SC1** | **6.25** | **9.82** |
| **SC2** | **7.45** | **9.63** |

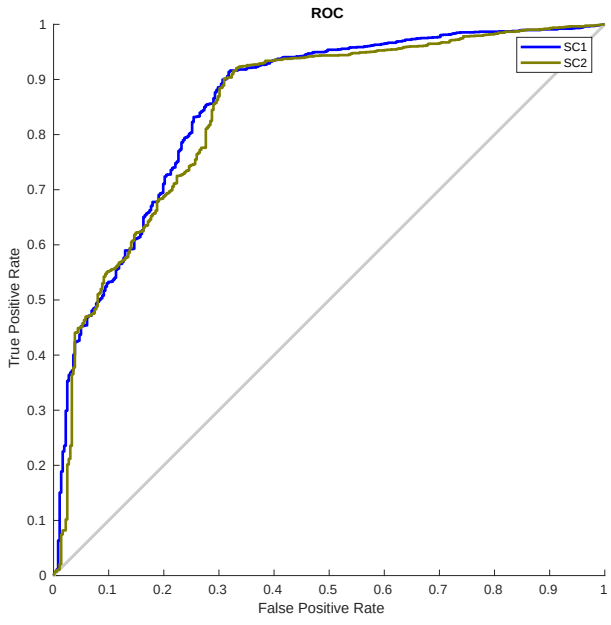


**Figure 8.** Receiver operating characteristic (ROC) curve of the proposed method. The blue and green curves are for the first and second scenarios, respectively. The Area Under the Curve (AUC) values for the two scenarios (SC1 and SC2) are 86% and 84%, respectively

Also, as discussed above about including false alarms (FA). One can see in Table 4 how the EER is decreasing considerably when the operator declares false alarms. In Folder 4, the EER decreases by

1.1% when the operator includes only one false alarm. Besides, in Folder 7 when the operator declares 5 false alarms, the EER decreases by 5.6%.

When dealing with the computational time, the comparison of our method with other state of the art methods is presented in Table 5, the results are only indicative and should be treated with caution. This is mainly due to the different hardware performances of each work. In our case, we used I7 CPU with 32 Gb of RAM and graphic card NVIDIA Quadro 2000 M. The proposed computational time presented in Table 5 is not optimal and does not reflect the full potential of our method. The main objective was to minimize the error and to propose a robust method for real-world application.

**Table 4.** Adaptation to false alarm detection.

| Folder | EERFL | EERPL | FA | EERFL | EERPL |
|---|---|---|---|---|---|
| 4 | 4.4 | 7.7 | 1 | 3.3 | 7.2 |
| 7 | 24.4 | 33.3 | 5 | 18.8 | 28.8 |

**Table 5.** Brief information about computational time of the proposed method (given in seconds per frame).

| Method | Boiman (IBC) [22] | Mahadevan (MDT) [43] | Roshtkhari (STC) [9] | Li [24] | Xiao [23] | Ours |
|---|---|---|---|---|---|---|
| Ped2 | 83 | 25 | 0.22 | 1.38 | 0.29 | **0.15** |

## 5. Conclusions

In this paper, a novel online adaptive method based on combination of pretrained 3D residual network and online classifier were developed and implemented. It is able to detect abnormal events, prevent the marginalization of normal behavior that rarely occurs during the training phase and adapt to the appearance of new normal events in the testing phase. In addition, our method does not require pretreatment methods such as tracking or background subtraction. This method is based on two main stages: Spatiotemporal feature extraction without any need of training, and the use of robust incremental classifier that prevents the redundancy of information in CCTV. It can also either be used online or offline.

We have tested our proposed methodology on two main datasets using crowded (Ped2) and non-crowded scenes (CapSec).

The results from the Ped2 dataset showed high performance in detection and localization for abnormal events based on The EERFL and EERPL. To the best of our knowledge, this method outperforms all existing techniques present in the literature and used for Ped2.

Besides, the fastness and the simplicity of this method allow us to use it for real-world application (case of CapSec).

The results presented in this paper showed the effectiveness of using this framework in detecting abnormal events. This method is robust, takes into account rare normal events present in the training phase. Besides, it can be incorporated in online CCTV. Moreover, the method can be adapted so that human operators select false alarms to prevent its future appearance, which is suitable for dynamic environment.

In this method, the localization of abnormal events is reflected as patches in the original image. In some cases, these patches may overflow on normal regions. In future work, we will focus on the detection at the pixel level to precisely localize the abnormal regions. Future studies will also investigate the test of this method on other datasets and will improve our local dataset (CapSec) in order to generate quantitative results and use it as a regular dataset for testing abnormal event detection methods. Moreover, we will also compare our classifier with other classification methods (k-nearest neighbors (k-NN) [47], and enhanced k-NN [48]).

**Author Contributions:** conceptualization, S.B. and H.S.; methodology, S.B. and H.S.; software, S.B.; analysis, S.B., H.S., M.M.H. and N.T.; data construction, S.B.; writing—original draft preparation, S.B. and N.T.; writing—review and editing, S.B., N.T., H.S. and T.W.; supervision—project administration—funding acquisition, H.S.

**Funding:** This work is supported by the French regional council of Grand-Est and the European regional development fund-FEDER.

**Acknowledgments:** The authors are grateful to anonymous reviewers for their comments that considerably enhanced the quality of the paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Sulman, N.; Sanocki, T.; Goldgof, D.; Kasturi, R. How effective is human video surveillance performance? In Proceedings of the IEEE 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–3.
2. Green, M.W. *The Appropriate and Effective Use of Security Technologies in US Schools: A Guide for Schools and Law Enforcement Agencies*; Technical Report; Sandia National Laboratories: Albuquerque, NM, USA, 2005.
3. Wu, S.; Moore, B.E.; Shah, M. Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2054–2060. [CrossRef]
4. Piciarelli, C.; Micheloni, C.; Foresti, G.L. Trajectory-based anomalous event detection. *IEEE Trans. Circ. Syst. Video Technol.* **2008**, *18*, 1544–1554. [CrossRef]
5. Jiang, F.; Yuan, J.; Tsaftaris, S.A.; Katsaggelos, A.K. Anomalous video event detection using spatiotemporal context. *Comput. Vis. Image Underst.* **2011**, *115*, 323–333. [CrossRef]
6. Ermis, E.B.; Saligrama, V.; Jodoin, P.M.; Konrad, J. Motion segmentation and abnormal behavior detection via behavior clustering. In Proceedings of the 15th IEEE International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008.
7. Reddy, V.; Sanderson, C.; Lovell, B.C. Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Colorado Springs, CO, USA, 20–25 June 2011; pp. 55–61.
8. Wang, T.; Snoussi, H. Detection of abnormal visual events via global optical flow orientation histogram. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 988–998. [CrossRef]
9. Roshtkhari, M.J.; Levine, M.D. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Comput. Vis. Image Underst.* **2013**, *117*, 1436–1452. [CrossRef]
10. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
11. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **2015**, *33*, 831. [CrossRef]
12. Heaton, J.; Polson, N.; Witte, J.H. Deep learning for finance: Deep portfolios. *Appl. Stoch. Mod. Bus. Ind.* **2017**, *33*, 3–12. [CrossRef]
13. Romero, A.; Ballas, N.; Kahou, S.; Chassang, A.; Gatta, C.; Bengio, Y. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Springer: Berlin, Germany, 2015.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
15. Zhou, S.; Shen, W.; Zeng, D.; Fang, M.; Wei, Y.; Zhang, Z. Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Process. Image Commun.* **2016**, *47*, 358–368. [CrossRef]
16. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 733–742.

17. Ravanbakhsh, M.; Nabi, M.; Sangineto, E.; Marcenaro, L.; Regazzoni, C.; Sebe, N. Abnormal event detection in videos using generative adversarial nets. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1577–1581.
18. Bouindour, S.; Hittawe, M.M.; Mahfouz, S.; Snoussi, H. Abnormal event detection using convolutional neural networks and 1-class SVM classifier. In Proceedings of the 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017), Madrid, Spain, 13–15 December 2017; pp. 1–6.
19. Calderara, S.; Heinemann, U.; Prati, A.; Cucchiara, R.; Tishby, N. Detecting anomalies in people's trajectories using spectral graph analysis. *Comput. Vis. Image Underst.* **2011**, *115*, 1099–1111. [CrossRef]
20. Morris, B.T.; Trivedi, M.M. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2287–2301. [CrossRef]
21. Antonakaki, P.; Kosmopoulos, D.; Perantonis, S.J. Detecting abnormal human behaviour using multiple cameras. *Signal Process.* **2009**, *89*, 1723–1738. [CrossRef]
22. Boiman, O.; Irani, M. Detecting irregularities in images and in video. *Int. J. Comput. Vis.* **2007**, *74*, 17–31. [CrossRef]
23. Xiao, T.; Zhang, C.; Zha, H. Learning to detect anomalies in surveillance video. *IEEE Signal Process. Lett.* **2015**, *22*, 1477–1481. [CrossRef]
24. Li, W.; Mahadevan, V.; Vasconcelos, N. Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 18–32. [PubMed]
25. Chong, Y.S.; Tay, Y.H. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*; Springer: Berlin, Germany, 2017; pp. 189–196.
26. Sabokrou, M.; Fayyaz, M.; Fathy, M.; Moayed, Z.; Klette, R. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Comput. Vis. Image Underst.* **2018**, *172*, 88–97. [CrossRef]
27. Sabokrou, M.; Fathy, M.; Hoseini, M.; Klette, R. Real-time anomaly detection and localization in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 56–62.
28. Xu, D.; Yan, Y.; Ricci, E.; Sebe, N. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.* **2017**, *156*, 117–127. [CrossRef]
29. Ravanbakhsh, M.; Nabi, M.; Mousavi, H.; Sangineto, E.; Sebe, N. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. *arXiv* **2016**, arXiv:1610.00307.
30. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 4489–4497.
31. Tran, D.; Ray, J.; Shou, Z.; Chang, S.F.; Paluri, M. Convnet architecture search for spatiotemporal feature learning. *arXiv* **2017**, arXiv:1708.05038.
32. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5534–5542.
33. Vapnik, V. Pattern recognition using generalized portrait method. *Autom. Remote Control* **1963**, *24*, 774–780.
34. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, 27. [CrossRef]
35. Burges, C.J. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167. [CrossRef]
36. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
37. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
38. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D.L. The mahalanobis distance. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18. [CrossRef]
39. Mehran, R.; Oyama, A.; Shah, M. Abnormal crowd behavior detection using social force model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, USA, 22–24 June 2009; pp. 935–942.

40. Adam, A.; Rivlin, E.; Shimshoni, I.; Reinitz, D. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 555–560. [CrossRef] [PubMed]
41. Bertini, M.; Del Bimbo, A.; Seidenari, L. Multi-scale and real-time non-parametric approach for anomaly detection and localization. *Comput. Vis. Image Underst.* **2012**, *116*, 320–329. [CrossRef]
42. Kim, J.; Grauman, K. Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, USA, 22–24 June 2009; pp. 2921–2928.
43. Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly detection in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1975–1981.
44. Sabokrou, M.; Fathy, M.; Hoseini, M. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electron. Lett.* **2016**, *52*, 1122–1124. [CrossRef]
45. Sabokrou, M.; Fayyaz, M.; Fathy, M.; Klette, R. Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Trans. Image Process.* **2017**, *26*, 1992–2004. [CrossRef]
46. Sabokrou, M.; Fathy, M.; Moayed, Z.; Klette, R. Fast and accurate detection and localization of abnormal behavior in crowded scenes. *Mach. Vis. Appl.* **2017**, *28*, 965–985. [CrossRef]
47. Fix, E.; Hodges J.L., Jr. *Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties*; Technical Report; University of California: Berkeley, CA, USA, 1951.
48. Nguyen, B.P.; Tay, W.L.; Chui, C.K. Robust Biometric Recognition From Palm Depth Images for Gloved Hands. *IEEE Trans. Hum.-Mach. Syst.* **2015**, *45*, 799–804. [CrossRef]