A Real-Time Deep Learning Approach for Real-World Video Anomaly Detection

Stefano Petrocchi
Department of Information Engineering, University of Pisa
Pisa, Italy
s.petrocchi@studenti.unipi.it

Giacomo Giorgi
Institute for Informatics and Telematics, National Research Council of Italy
Pisa, Italy
giacomo.giorgi@iit.cnr.it

Mario G. C. A. Cimino
Department of Information Engineering, University of Pisa
Pisa, Italy
mario.cimino@unipi.it

ABSTRACT

Anomaly detection in video streams with imbalanced data and real-time constraints is a challenging task of computer vision. This paper proposes a novel real-time approach for real-world video anomaly detection exploiting a supervised learning methodology. In particular, we present a deep learning architecture based on the analysis of contextual, spatial, and motion information extracted from the video. A data balancing strategy based on hard-mining and adaptive framerate is used to avoid overfitting and increase detection accuracy. The approach defines an extended taxonomy by differentiating anomalies in "soft" and "hard". A novel anomaly detection score based on a sigmoidal function has been introduced to reduce false positive rate while maintaining a high level of true positive rate. The proposed methodology has been validated with a set of experiments on a well-known video anomaly dataset: UCF-CRIME. The experiments on the testbed demonstrate the impact of the contextual information and data balancing on the classification performances, considering only "hard" anomalies during training and that the proposed model can achieve state-of-the-art performances while minimizing resource consumption.

CCS CONCEPTS

• **Computing methodologies** → **Scene anomaly detection**; • **Applied computing** → *Surveillance mechanisms.*

KEYWORDS

Anomaly detection, Behavioral analysis, Deep Learning, Computer vision

1 INTRODUCTION

Nowadays, the advances in Information and Communication Technologies (ICT) have led to the transformation of the environments in intelligent entities, e.g., smart-home, smart-buildings, smart-city that offer a set of interconnected devices empowered by computational and wireless communication capabilities. Such devices can provide services to everyday activities for better quality living. Network cameras are widely used in such environments and are exploited to develop applications such as license plate recognition [37], people counting [5], vehicle/person tracking [2, 35], safety control of smart objects [11]. In the last years, the public safety aspects have increasingly gained attention in our society. Many applications in the video surveillance field have been developed through the diffusion of networks of cameras and the advances in Artificial Intelligence (AI). One of the main tasks in the video surveillance field is given by the supervision of multiple monitors to the end to detect anomalous behavior as quickly as possible. Therefore, there is a growing demand for an intelligent video surveillance system that automatically and real-time detects anomalous behavior, e.g., crimes, illegal activities, or environmental incidents, and timely raises the alarm. In most cases, anomalous behaviors are rare events that do not occur frequently or rapid behaviors that happen in few seconds. Therefore, they become difficult to capture. Moreover, video information is challenging to represent due to its high dimensionality and noise that can affect the scene, e.g., occlusion, low camera quality, high brightness. In addition to these factors, anomaly behaviors are, by their nature, highly correlated to the context, i.e., the running behavior is in itself normal but can be considered anomalous if it is done in an inappropriate context. These challenges have made it difficult for machine learning methods to identify video patterns that produce anomalies in real-world applications. There are many successful cases in the related field of behavioral recognition [17, 27]. However, these methods detect behaviors that consist of clearly defined actions in pre-defined contexts, without significant video noise issues. Other works that treat the task as a binary classification problem (anomalous or normal behavior) [25] proved to be accurate to detect anomalies. However, they are validated in datasets where the anomalies are related to limited contexts and can be easy to distinguish, i.e., a moving car in a pedestrian street [29].

The rest of the paper is structured as follows. Section 2, reports different literature works related to the anomaly detection problem and highlights their advantages and disadvantages. Section 3

presents the proposed anomaly detection's methodology. In Section 4 we describe the dataset used and the experiments conducted. Section 5 shows the results obtained from the experiments and provides a discussion on the proposed methodology's impact. Section 6 briefly concludes and proposes also some future work directions. To summarize, this paper provides the following contributions:

- We propose a novel lightweight deep learning architecture for real-time video anomaly detection that considers contextual, spatial, and motion information.
- We introduce a novel anomaly score mechanism based on the sigmoidal function that produces a more robust anomaly score reducing the false-positive rate, maintaining a high level of true positive rate.
- We propose a data balancing mechanisms to be applied during training and based on hard-mining and an adaptive sampling rate.
- We defined two types of anomaly frames based on the temporal evolution of the anomalous activity. *Soft anomaly frames*, i.e., frames in which something anomalous is going to happen or that happened, without the anomaly being properly in progress, and *hard anomaly frames*, i.e., frames in which the anomaly is currently in action.
- We analyzed the learning activity using the GradCam explainability tool to assess which part in the frame contributes to the anomaly detection and exploit the results as a tool for anomaly localization in the scene.

Link to code:
github.com/iitcybersecurity/RealWorldVideoAnomalyDetection

2 RELATED WORKS

Video anomaly detection is one of the most complexes and studied problems of computer vision. An *anomaly* is any pattern that does not conform to what is considered *normal* [1]. The two classes that characterize the problem are defined as one the negation of the other and both can take completely different forms, specific to the problem under consideration, i.e., a person who runs can be considered an anomaly within a public office but is normal in contexts such as stations or parks. Moreover, it is not easy to establish a priori all the possible anomalies that may occur even considering a single context.

Handcrafted Methods. The first methods of anomaly detection were *trajectory-based* [19, 23]. The main idea is to identify the distinct trajectories of objects within normal videos. The anomalies are highlighted as objects that do not follow similar trajectories. However, these methods are of restricted applicability and can be used only in the presence of constant and unobstructed trajectories. The use of other handcrafted features allows enriching the ability of a detector to identify more general classes of anomalies, not limited to trajectories only. These low-level features generally extrapolate information about appearance, movement, and texture. *Histograms of optical flows* [7], *histograms of oriented gradients* [18], *social forces maps* [30] and *mixture of dynamic textures* [24], are just some of the methods developed. Although very effective in identifying specific anomalies, these feature extraction methods cannot adapt to categories of abnormalities not previously seen.

Semi-Supervised Methods. To overcome these problems, some of the most used approaches are typically *semi-supervised*. This learning method category uses only normal videos to train the detector to identify anomalies as any deviation from the notion of normality that they have learned. Precisely avoiding giving a specific characterization to anomalies allows this type of detector to be more robust towards types of abnormalities not initially foreseen. Another great help given to the generalization capabilities of anomaly detectors is the use of features extracted through *Deep Neural Networks* (DNNs) [21]. Neural networks allow to autonomously extract semantically significant features that can introduce a better generalization capability with respect to the handcrafted ones. The current state-of-the-art combines semi-supervised approaches and neural networks. Among the most popular approaches in literature, we can mention *autoencoders* [13] and *Generative Adversarial Networks* (GAN) [31]. These networks are usually trained on images – frame and optical flow – extracted from normal videos to reconstruct them or predict the next in time order [28]. When anomalous images are presented to the network this is generally not able to recreate them, as it is trained only on normal images. The anomaly generates a greater reconstruction error, allowing it to be distinguished. Although certainly more robust than handcrafted methods, techniques based on image reconstruction also have limitations, i.e., the networks may be able to reconstruct even the anomalies [12], not allowing them to be distinguished. In [3] it is highlighted that deep learning method are also characterized by a lack of explainability. In this work, the GradCam tool [32] is used as a method to locate the regions in a frame that contributes most to the assignment of a higher reconstruction error in their auto-encoder based approach.

Supervised Methods. These approaches involve the use of labeled videos to reduce the problem of binary classification. The main obstacle to studying these methods is the non-availability of large labeled datasets, which are very expensive to produce. Recently [34] introduced a new dataset for the detection of real-life anomalies concerning crimes – e.g., robberies, assaults, shootings – with more than 128 hours of untrimmed videos (*UCF-Crime*). The same paper proposes a *multiple instance learning* method that allows the training of *weakly supervised* binary classifiers using labels at video level. More recently in [22] the UCF-Crime dataset has been enriched with spatiotemporal annotations (*UCFCcrime2local*), allowing the experimentation of *strongly supervised* methods. Some commonly used neural networks in a supervised environment are *3D convolutional networks* [36]. These are also often exploited in the field of *action recognition* [38] and allow to extrapolate spatiotemporal features able to describe the actions inside video segments. They can also be used in the form of *two-stream networks* [6] to extract features from streams with different frame rates or using an optical flow stream in parallel with the video stream. However, 3D networks have the problem of being particularly heavy, both for training and inference. For practical applications, lighter networks such as *2D-CNNs* can be taken into consideration by enriching their capabilities using solutions to add time and motion information to the spatial features extracted from the network. Temporal

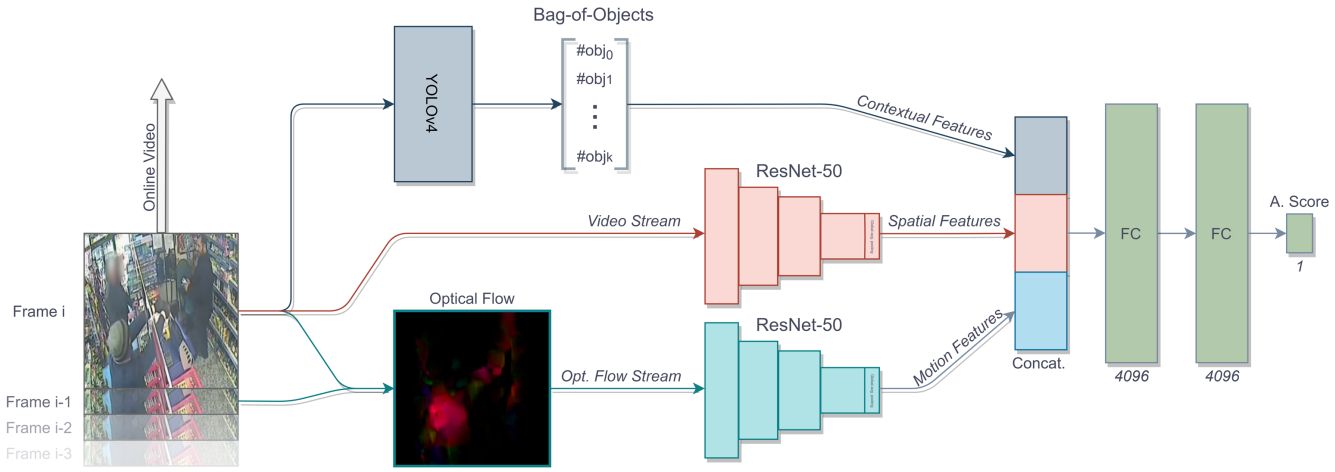

Figure 1: Diagram of the proposed model. The double stream network consists of two ResNet-50 that take in input the current frame and the optical flow calculated between current and previous frames. The features extracted from the double stream network are concatenated to the bag-of-objects made using YOLOv4 object detector on the current frame. The concatenated features are then forwarded to two fully connected layers of 4096 neurons to produce the classifier anomaly score.

information can be added, for example, by feeding *Long-Short-Term-Memory networks* (LSTM) with the spatial features extracted by the CNN as in [15] while motion information can be included using a two-stream solution [33].

3 PROPOSED METHODOLOGY

The following section describes the proposed model together with techniques for efficient and effective data utilization during training and a novel approach for anomaly score.

3.1 Contextual Information Extraction

The object detector used is *YOLOv4* [4]. It is currently state-of-the-art in terms of efficiency and accuracy while still having a rate of inference that allows real-time video analysis. We use a version of YOLO pre-trained on *MS COCO* [26], a dataset with 80 classes that include people, animals, vehicles, and everyday objects. Video frames are given in input to YOLO to create *contextual features* used by a classifier downstream. Such features are constructed as a vector containing the number of objects identified for each class in the frame, which from now on will be defined as *bag-of-objects* (see Figure 1). The idea is that in a context such as that of UCF-Crime, the introduction of information on specific classes of objects in the scene may be useful to a classifier that can weigh these features against other factors such as temporal and spatial information.

3.2 Spatial and Motion information extraction

Spatial and motion features are extracted through a two-stream architecture. Figure 1 shows the proposed model that involves a pretrained *ResNet-50* [14] on *ImageNet* [8] as base convolutional network. In particular, the model receives as input streams the frames and optical flow (pattern of apparent motion caused by the relative movement between an observer and a scene [16]) extracted from the video that is being analyzed. The features obtained from the two streams are extracted from the last convolutional layer through a global average pooling. These features, according to a *joint fusion* architecture, are concatenated to the bag-of-objects extracted from the current frame and given in input to a fully connected classifier. Specifically, this is composed of two fully connected layers with *ReLu* activation, consisting of 4096 neurons each and connected to a final neuron that produces as anomaly score the *anomaly class confidence* between $[0, 1]$.

3.3 Data Balancing

One of the main issues in the training of the anomaly detection model lies in the balance between the classes. It concerns both the higher absolute number of frames considered normal compared to those containing anomalies and their distribution in the videos. A supervised model needs to use a training distribution of videos that include as many different types of anomalies as possible; in addition, for the same anomaly (e.g., shooting, robbery), it is important to have multiple different videos in different scenarios contexts. The solution proposed to manage these issues involves using different sampling framerates for anomalous and normal segments and hard mining.

Adaptive Sampling Framerates. During the training phase, it is necessary to subdivide the videos into single frames or segments of adjacent frames respectively for 2D-ConvnNets and 3D-ConvNets. For 2D-CNNs, the most straightforward solution is to split videos into single frames respecting the original framerate, for instance, 30 FPS in UCF-Crime. Applying this mechanism, we obtained a large amount of redundant data: two adjacent frames contain only slight differences that require an increasing training cost and can produce overfitting problems. A simple solution to this problem is using a lower framerate for frame sampling than the original one. In this way, with the same resources, it is possible to use a greater variety of videos to allow the network to generalize the anomalies better. This method can also be used in an adaptive

way to sample with an even lower framerate normal videos. In such a context, sudden movements are not relevant. Thus an undersampling does not reduce the training quality. The same principle can be applied in 3D-CNNs where the stride (the overlap between two segments extracted consecutively) can be greater for segments containing anomalies and lower for normal ones.

Hard Mining. As described in [9] hard mining is a method that involves selecting – mine – the samples with greater loss to be used during the training phase. In our approach is used a batch-wise hard mining. Considering using batches of K samples during training, for each batch, we sample without replacement αK candidates frames, where α is a multiplicative factor. First, an inference operation is performed on the candidates to calculate the losses; after, only the K first samples with greater loss are selected, regardless of whether they are hard-positives or hard-negatives, and used as a real batch training. In such a way, in each epoch, only $\frac{1}{\alpha}$ of the total training samples are used. Hard mining impacts anomaly detection problems because they are characterized by a large variety, although these consist of only two classes. In particular, the anomalous class is actually made up of several sub-classes, the individual anomalies which are equally important and must all be recognized by a classifier. Hard mining, therefore, allows for adaptive sampling of these sub-classes in every epoch.

3.4 Hard and Soft Anomalies

In untrimmed videos, the distinction between anomalous video segments and normal segments may become unclear. Before an anomaly and immediately after its end, we can consider video segments in which it is respectively guessable that something anomalous will happen or that happened, without the anomaly being properly in progress. For example, before a theft of objects in a car is possible to guess that something will happen if someone suspicious is looking through the windows of the vehicle. In the same way, it is guessable that a theft occurred if a car has a broken glass and it is scattered on the ground. These suspicious video segments can be considered as *soft anomalies*. They are not properly normal segments as they have ambiguous behaviors in them. However, they cannot be considered even true anomalies (*hard anomalies*) as they do not contain real crimes. Labelling soft anomalies as completely normal or completely anomalous should therefore be avoided in order to improve training quality.

3.5 Sigmoidal Anomaly Score

In [10] the classifier's output value is not used directly to perform online anomaly detection, but they preferred to use a more noise-robust method to calculate the final *anomaly score*. Similarly in our model we could use the confidence value of the anomalous class as anomaly score and alert when this value is higher than a certain alarm threshold. The score thus computed, however, is very noisy and causes frequent false positives. For this reason, we propose the ***sigmoidal anomaly score*** (SAS) as novel method for calculating the anomaly score using only two main parameters: *sensibility* and *reactivity*. Anomaly score s_t is calculated as:

$$s_t = S(x_t) = \frac{1}{1 + e^{-x_t}}$$

Where $S : \mathbb{R} \rightarrow [0, 1]$ is a standard logistic sigmoid function and x_t the value of the *accumulator* at time t computed as:

$$x_t = \begin{cases} x_{t-1} + \Delta_t^+ \nu & \text{if } \sigma_t \geq \tau \\ x_{t-1} - \Delta_t^- \nu & \text{if } \sigma_t < \tau \end{cases} \quad with\ x_t \in [LB, UB]$$

Where σ_t is the anomaly class confidence calculated at time t by the classifier, x_{t-1} is the accumulator's value calculated at previous step, $\tau \in [0, 1)$ is the sensibility threshold and $\nu \in (0, +\infty)$ the reactivity parameter. $\Delta_t^+, \Delta_t^- \in [0, 1]$ can be determined as:

$$\Delta_t^+ = \frac{\sigma_t - \tau}{1 - \tau};\ \Delta_t^- = \frac{1 - \sigma_t - \tau}{1 - \tau}$$

They are the increase or decrease in score caused by exceeding or not the sensibility threshold τ. The closer τ gets to 1, more confidence is needed in anomaly classification to increase the score value, the closer τ is to 0, more sensitive the score is to lower confidence classifications. Reactivity ν is instead a multiplicative factor that allows to decide how quickly the score grows or decreases: a high value of ν will make the score more sensitive to anomalies that take place in short time intervals, but making false positives due to noise more likely. A value of ν below 1, on the contrary, allows to filter the noise more effectively and makes the score less sensitive to sudden anomalies. To avoid that the accumulator x assumes too high or low values, these are limited by a lower and upper bound (LB, UB).

4 EXPERIMENTS

This section describes in detail all the experiments carried out to validate the proposed model and methodologies.

4.1 Dataset

In our experiments, we used the UCF-Crime dataset developed by [34]. It consists of long untrimmed surveillance videos which cover 13 real-world anomalies: *Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting,* and *Vandalism*. The original annotations of the dataset are at video level. To our scope, we used the frame-level annotations provided by [22] in UCFCrime2local. They created a training set of 210 videos and a test set of 90 videos with 7 of the 14 original classes: *Arrest, Assault, Burglary, Robbery, Stealing, Vandalism* and *Normal*. The annotations are in *Vatic* format and we decided to consider as *hard anomalies* the frames with *lost flags* equal to 0 and the others as *soft anomalies* because the anomaly is not clearly present. All frames in normal videos were considered normal.

4.2 Impact of Bag-of-Objects on Anomaly Detection

The purpose of this experiment is to demonstrate how the contextual information provided by the bag-of-objects have a positive impact on the detection of anomalies. Four different models have been trained in these experiments: a 2D-two-stream-CNN with the use of bag-of-objects (***2S-bag***), a 2D-two-stream-CNN without the use of bag-of-objects (***2S-no-bag***), a single flow CNN (only frames) with the use of bag-of-objects (***1S-bag***) and a single flow CNN without the use of bag-of-objects (***1S-no-bag***).

Training Details. As preprocessing, the images are resized to 224x224 and converted from RGB to BGR, then each color channel is zero-centered with respect to the ImageNet dataset, without scaling. Half of the test set videos were divided to create the validation set to preserve the class distribution. The validation set is used together with the early stopping technique to keep the overfitting under control. The sampling framerate was set to 3 FPS for video segments containing anomalies and 10 FPS for normal ones. Segments without hard anomalies were considered normal. Hard mining with $\alpha = 3$ has been used, and moreover, class weights were set to 1 for normal class and 2 for anomaly class. *Adam* is used as optimizer, and for each model were initially trained only the final fully-connected layers using a learning rate of 10^{-5} until reaching the lowest loss on validation set with the patience of 5 epochs. Then the networks were fully trained using the same methodology but with a learning rate of 10^{-6}. After each fully connected layer was applied a dropout of 0.5.

4.3 Impact of Hard Mining on Anomaly Detection

The second experiment demonstrates the effectiveness of the hard mining technique even on binary classification problems. Two models were compared: a 2D-two-stream-CNN with the use of hard mining (the same *2S-bag* than before) and a version trained without its use (***2S-bag-no-hm***). The remaining training parameters have been left unchanged. The comparison's fairness is guaranteed by using a very low learning rate and early stopping, allowing the networks to achieve optimal accuracy. Indeed, it must be considered that with hard mining, an epoch contains $\frac{1}{\alpha}$ of the samples of a training epoch that do not exploit it.

4.4 Impact of Soft Anomalies on Anomaly Detection

In this experiment, we want to demonstrate how the use of the most ambiguous frames, which we have identified as soft anomalies, significantly impacts the model's performance during training. Following the same training guidelines of the experiment described in the previous section, we compared the *2S-bag* and *1S-bag* models trained with the difference of not considering those frames in anomalous videos that do not contain an anomaly (hard anomalies), neither as normal frames nor anomalous frames, but simply excluding them. We will refer to the models trained with only hard anomalies as ***2S-bag-ha*** and ***1S-bag-ha***. In order to have a fair comparison with the previous experiments, no frames were excluded from the test set, and those with soft anomalies were considered normal.

4.5 Proposed Methodologies Applicability to Alternative Models

To verify the versatility of the techniques described above, these have also been tested on a 3D-CNN alternative (***C3D-ha***). 3D neural networks allow the extraction of spatiotemporal features taking in input segments of multiple contiguous frames. The model we tested is the *C3D* developed by [36] and pre-trained by the same authors on *Sport1M* [20], an action recognition dataset containing one million clips of different sports. Spatiotemporal features are extracted from the last convolutional layer through a flattening operation. These are then given in input to a fully connected classifier analogous to the one used in Figure 1.

Training Details. As preprocessing step was only performed a resizing of the frames to 112x112. The network takes in input segments of 16 frames. These were sampled from the original videos at 7.5 FPS (1/4 of the original framerate) to create segments of about 2 seconds like those used for pretraining. The concept of the different sampling rates for anomalous and normal sections was interpreted in this case using different strides: anomalous segments are sampled with an overlap of 12 frames (a shift of 4), while normal segments are taken with an overlap of 8 frames (a shift of 8). The training is carried out in the same way as previously described using hard mining, early stopping, a two-phase fine-tuning, and considering only hard anomalies. The only difference is that segments are used instead of frames as unitary samples.

4.6 Comparison with the State of the Art

Our best model (*2S-bag-ha*) has been compared with the results obtained by [22]. The architecture they tested is a two-stream I3D with RGB and optical flow volumes in input. In particular, in their work, a comparison was made between (i) a model trained using full-frames (***full-I3D***), (ii) a strongly supervised model trained with ground-truth spatial annotations (***oracle-I3D***) and (iii) a weakly supervised model (***ws-I3D***) trained using annotations made exploiting the strongly supervised one. To make the comparison fair, we retrained our model without using the early stopping, but fixing 5 epochs for the first stage of training and 2 for the second (***2S-bag-ha-no-es***). In this way, it was possible to use exactly the same split indicated in their work both in training and in testing.

4.7 Generalization Capabilities

The main objection that can be made to the models developed in this work is the strongly supervised nature of the training. It is possible to think that the models effectively recognize only the classes of anomalies provided in training and that their performance is much lower on classes never seen before. For this reason we verified the generalization capabilities of the *2S-bag-ha* model on the remaining UCF-Crime classes that were not labeled in UCFCrime2local and thus not used in training: *Shoplifting*, *RoadAccident*, *Shooting*, *Fighting*, *Abuse*, *Arson* and *Explosion*. Although coming from the same dataset, some of these classes of anomalies are fundamentally different from those used in training and is not trivial that the model could correctly recognize them as well. Therefore, it was constructed a new test set containing the 450 anomalous videos of the unlabeled classes and 150 new normal videos, all not used during training.

5 EXPERIMENTAL RESULTS

In this section we discuss the results of the experiments previously described.

Model	FL-AUC (%)	VL-F_1 (%)
2S-bag	82.6	66.7
2S-no-bag	77.6	61.5
2S-bag-no-hm	80.2	64.0
2S-bag-ha	**83.7**	**91.0**
1S-bag	78.6	74.1
1S-no-bag	75.3	58.3
1S-bag-ha	**81.5**	**87.5**
C3D-ha	**83.7**	69.0

Table 1: Experimental Results

Model	FL-AUC (%)
full-I3D	56.1
oracle-I3D	74.7
ws-I3D	81.0
2S-bag-ha-no-es	**82.4**

Table 2: Comparison with the state-of-the-art

Model	VL-F_1(%)	VL-TPR(%)	VL-FPR(%)
2S-bag-ha	89.7	86.4	18.6

Table 3: Generalization Capabilities

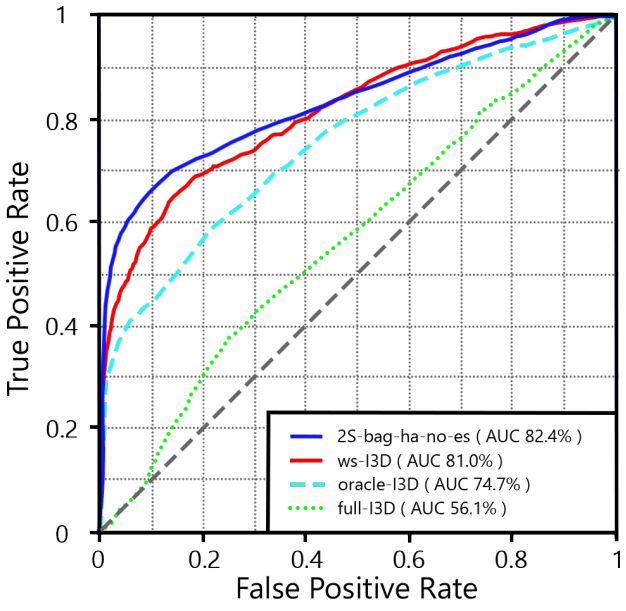

Figure 2: ROC comparison between proposed *2S-bag-ha-no-es* model (blue) and *ws-I3D* (red), *oracle-I3D* (dashed cyan) and *full-I3D* (dotted green) models presented in [22].

5.1 Evaluation metrics

As in [34] and [22] the various models were compared using the ***area under the ROC curve*** calculated frame by frame (FL-AUC). In the case of the 3D-CNN model, the classification is considered on the last frame of the segment as the model is intended to be used online. It was also introduced a new metric to evaluate the effectiveness of the models in the classification at the video-level: the ***video-level F_1 score*** (VL-F_1). This metric is useful to evaluate the models' anomaly detection capabilities with a weakly-labeled test set. The same model can obtain different results for the two metrics. For example, it can distinguish normal videos from anomalous ones correctly but not normal and anomalous segments inside the same video. It is therefore important to use both metrics to have a correct comparison. Using the VL-F_1 score, a whole video is classified as anomalous if its *sigmoidal anomaly score* exceeds 0.5 at any point and normal otherwise. Once all videos have been classified, the F_1 score is calculated as:

$$VL\text{-}F_1 = \frac{2}{\frac{1}{vl\text{-}r} + \frac{1}{vl\text{-}p}}$$

Where *vl-r* is the *video-level recall* and *vl-p* the *video-level precision*. SAS parameters have been heuristically set with: $\tau = 0.5$, $\nu = 2$, $UB = 7$ and $LB = -7$.

5.2 Evaluation of Proposed Techniques Impact on Anomaly Detection

The results of the methodologies used to enhance the model's accuracy and training's efficiency are described below.

Bag-of-Objects. Referring to Table 1 is possible to notice how the use of bag-of-objects has a clear impact on the model's ability to classify at frame-level. In particular, both two-stream and one-stream architectures have an AUC increase of 6.4% and 4.4%: from 77.6% (2S-no-bag) to 82.6% (2S-bag) and from 75.3% (1S-no-bag) to 78.6% (1S-bag). Also, with regard to the video-level F_1 score both architectures have an increase of 8.5% and a remarkable 27.1%: from 61.5% (2S-no-bag) to 66.7% (2S-bag) and from 58.3% (1S-no-bag) to 74.1% (1S-bag). It certifies that the use of the bag-of-object has a positive impact on the classification of the single frames and the distinction of videos containing anomalies from normal ones.

Hard Mining. Always referring to Table 1 the impact of hard mining on anomalous detection can be verified by comparing 2S-bag with 2S-bag-no-hm. There is an increase in AUC accuracy of 3%, from 80.2% to 82.6%, with the use of hard mining and also the VL-F_1 score increases by 4.2%, from 64.0% to 66.7%.

Soft Anomalies. In Table 1 can also be assessed the impact of the exclusive use of *hard anomalies* as positive samples. For both architectures the maximum AUC values are reached, with an increase of 1.3% and 3.7%: from 82.6% (2S-bag) to 83.7% (2S-bag-ha) and from 78.6% (1S-bag) to 81.5% (1S-bag-ha). But the most significant results are the VL-F_1 values with a significant increase of 36.4% and 18.1%: from 66.7% (2S-bag) to 91.0% (2S-bag-ha) and from 74.1% (1S-bag) to 87.5% (1S-bag-ha). It means that an accurate selection of abnormal and normal frames – so that they are completely unambiguous

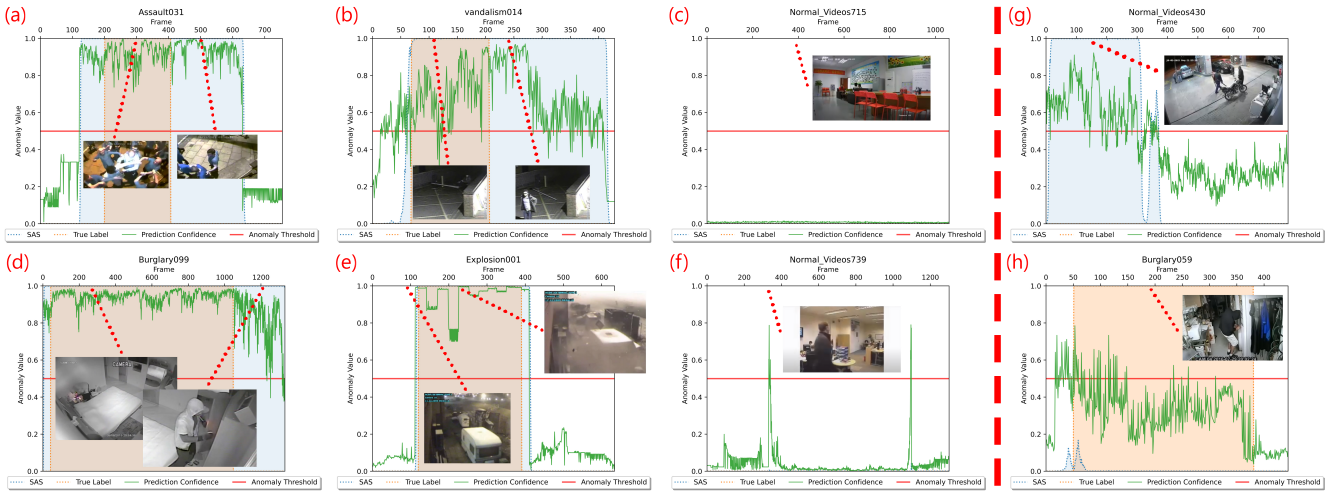

Figure 3: Qualitative results for our *2S-bag-ha* model on testing videos. Orange frame windows indicate a region labeled as anomalous, light blue regions show sigmoidal anomaly score (SAS) values, the green line shows classifier's confidence, and the red line shows the threshold above which the value of SAS signals the video as anomalous. (a), (b), (d) show anomalous videos where different anomalies are fully identified by our method. (e) shows a class of anomaly not used during training but correctly detected. (c) shows the correct response of our method to a normal video. (f) shows noise in the classification that is correctly filtered by the SAS method. (g) and (h) show a false alarm and a false negative case.

for their use in training – leads to a general improvement of the frame-level classification and allows a considerably better video-level detection.

Alternative Models. In the last row of Table 1 are present the results obtained for the *C3D* model. It achieves the best frame-level AUC at 83.7% tied with *2S-bag-ha*, but a noticeable lower value of VL-F_1 with 69%. It indicates that the various techniques presented in this work can also be successfully applied to other types of architectures and that temporal information may not be as important in video-level anomaly detection to distinguish soft and hard anomalies inside an untrimmed video, opening space for future studies.

5.3 Comparative Results with the State of the Art

Table 2 and Figure 2 involve a comparison between our model and those presented in [22] performed using the same dataset split and their annotations. It is possible to notice that *2S-bag-ha-noes* has a frame-level AUC value of 1.7% greater than their best model. The comparison becomes even more clear if we consider that our model uses the entire frame, while the *oracle-I3D* and *ws-I3D* models need spatial annotations to exploit the spatiotemporal tubes locality. Moreover, our model exploits during training only $\frac{1}{3}$ of the anomalous frames and less than $\frac{1}{10}$ of the normal frames available in training set by using the proposed adaptive under-sampling method. Therefore, it becomes clear how it reduces drastically the data needed during training and increases the final accuracy since our model trained on full-frames has a remarkable AUC increase of 46.9% compared to their model trained in the same way (full-I3D).

5.4 Generalisation Capabilities Evaluation

Referring to Table 3 are used with the video-level F_1 score the values of the video-level *true positive rate* (VL-TPR) and *false positive rate* (VL-FPR) obtained by *2S-bag-ha* on a dataset composed of 450 videos of anomalous classes never seen before by the network plus other 150 normal videos as negative examples. The network gets a good value of VL-F_1 at 89.7%, confirming an excellent generalization capability with a VL-TPR at 86.4%. It is also possible to notice that the VL-FPR values are contained at 18.6%. These results demonstrate that the network maintains almost the same levels of accuracy for unseen anomalous sub-classes as in the identification of training anomalies. Therefore, the network was successfully able to learn a general representation of the problem. Finally, all this result does not affect the models capacity to construct a generalized internal representation of the problem even in a highly supervised scenario.

5.5 Qualitative Results

In this subsection are presented some qualitative results obtained with the proposed model in order to evaluate its possible use in security applications.

Anomalous Activity Recognition. Figure 3 shows *2S-bag-ha* model results for anomalous activity recognition in testing videos. Concerning *hard and soft anomalies*, figures (a), (b), and (d) are some examples of anomalous videos in which the anomalous activity has been correctly identified. It is possible to notice the tendency of the anomaly detector to highlight as anomalous an area greater than that described as properly anomalous in the ground truth. This can be seen as another evidence of the importance of distinguishing

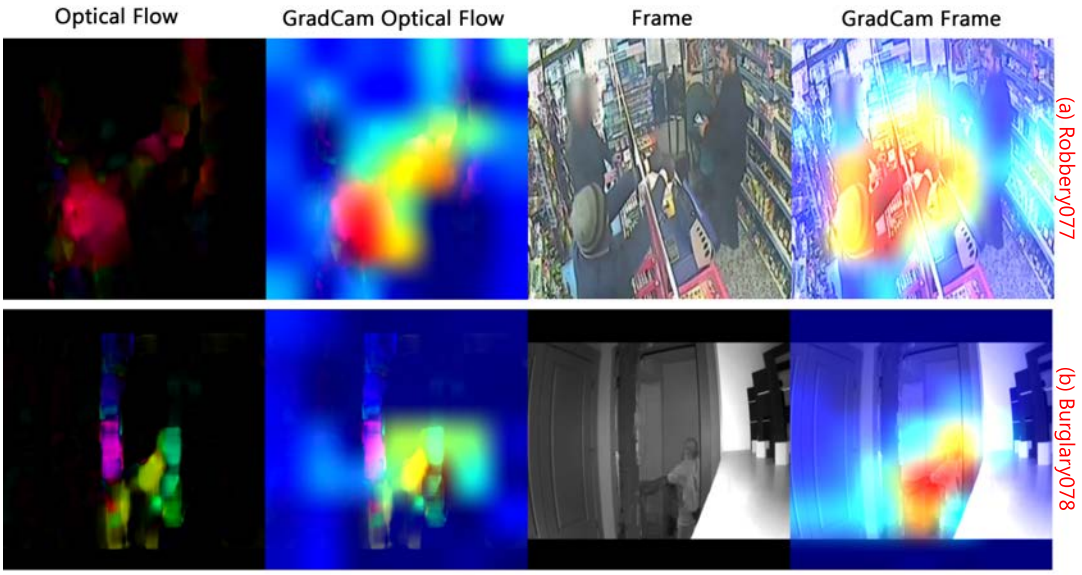

Figure 4: GradCam extracted from *2S-bag-ha* streams last convolutional layers and applied to two anomalous test videos. From right to left the images represent the optical flow calculated taking into account the apparent movement between previous and next frame, the GradCam applied on the optical flow, the relative frame extracted from the video and the GradCam applied to it.

soft anomalies from both hard anomalies and normal video sections. In (a) and (b), the sections that in the ground truth have not been labeled as anomalous (the crime is no longer visible) but are still reported as anomalous by our method, falls within the definition of soft anomalies. In particular, in (a), the second frame does not contain an assault but still includes an anomalous grouping of peoples. In (b), instead, the second frame shows a man leaving after vandalizing a bar that is now on the ground. Therefore the consequences of the vandalism are visible even if this is no longer ongoing. Regarding the analysis of the normal Activities and the SAS Noise Filtering, in (c) e (f), it is possible to notice how our SAS method correctly gives at videos that contain purely normal events an almost zero score. The effectiveness of this method is particularly evident when considering situations like the one in (f) in which can filter out momentarily incorrect classifications. These can happen because our model classifies considering individual frames, and a particular noise or position can produce an incorrect classification for the single frame. Thus the SAS helps in reducing the number of false positives that would be present using only a threshold on the classifier's confidence while maintaining a high true positive rate.

GradCam. For security applications, it is important to verify the video regions that contribute most to anomaly detection to assess the network's decision-making process and ensure that it has no biases. For this purpose, the *GradCam* method introduced by [32] allows identifying within an image is input to a ConvNet the regions that contribute most to the classification of a certain class. Identifying these regions can also have practical purposes to locate anomalies in progress, for instance, if the video takes a wide scene. In Figure 4 the GradCam method has been applied to *2S-bag-ha*. The simplicity that characterizes our model makes it possible to identify what the network interprets as anomalous, even *in real-time*. For example, both in case (a) and (b), the GradCam highlight the arms stretching and the resulting optical flow as major contributors to the detection. This gesture is common in almost all classes of anomalies on which the network has been trained. We can therefore think of combining the SAS and GradCam to create a surveillance system that allows not only to alarm promptly in case of an anomaly but also to highlight in real-time its position within the video.

6 CONCLUSION AND FUTURE WORKS

This paper has proposed a deep learning approach for the real-time detection of real-world behavioral anomalies in a video surveillance scenario. The methodology presented deals with the complexities and challenges of video anomaly detection: (i) the noisy and low-quality video stream, (ii) the high computational overhead introduced by the systems, and (iii) the difficulty in recognizing if action is anomalous depending on the context. To this end, our methodology proposes the addition of contextual information to spatial and motion features. We also introduce new data balancing strategies and the definition of a novel sigmoidal anomaly score. The combined use of those methodologies contributes to increasing the anomaly detection accuracy while reducing the resources used in training. Furthermore, the experimental results on the UCF-CRIME dataset show how the proposed method can achieve state-of-the-art performance by using a more lightweight architecture. The proposed approach has a wide margin for improvements in future

works. In fact, considering very generic classes, the contextual information given by the bag-of-object can be enriched considering also more relevant objects for the specific anomaly detection task. Moreover, the adaptive sampling rate technique can be made even more effective if, for example, the adaptivity also considers the different types of anomalies and their characteristics. Finally, the model could be trained to distinguish soft anomalies as a different *suspicious* class, using for instance, a fuzzy assignment.

ACKNOWLEDGMENTS

This work has been partially funded by EU funded project H2020 SIFIS-Home GA ID:952652.

REFERENCES

[1] 2020. A Survey on Deep Learning Techniques for Video Anomaly Detection. *arXiv preprint arXiv: 2009.14146* (2020).

[2] Ejaz Ahmed, Michael Jones, and Tim K Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3908–3916.

[3] Sukalyan Bhakat and Ganesh Ramakrishnan. 2019. Anomaly Detection in Surveillance Videos. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* (Kolkata, India) *(CoDS-COMAD '19)*. Association for Computing Machinery, New York, NY, USA, 252–255. https://doi.org/10.1145/3297001.3297034

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

[5] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. 2016. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 24th ACM international conference on Multimedia*. 640–644.

[6] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[7] Rensso Victor Hugo Mora Colque, Carlos Caetano, Matheus Toledo Lustosa de Andrade, and William Robson Schwartz. 2016. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 3 (2016), 673–682.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[9] Qi Dong, Shaogang Gong, and Xiatian Zhu. 2017. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 1851–1860.

[10] Keval Doshi and Yasin Yilmaz. 2021. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition* 114 (2021), 107865.

[11] Giacomo Giorgi, Antonio La Marra, Fabio Martinelli, Paolo Mori, and Andrea Saracino. 2017. Smart parental advisory: A usage control and deep learning-based framework for dynamic parental control on smart TV. In *International Workshop on Security and Trust Management*. Springer, 118–133.

[12] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1705–1714.

[13] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 733–742.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Jordan Henrio and Tomoharu Nakashima. 2018. Anomaly Detection in Videos Recorded by Drones in a Surveillance Context. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2503–2508.

[16] Berthold KP Horn and Brian G Schunck. 1981. Determining optical flow. *Artificial intelligence* 17, 1-3 (1981), 185–203.

[17] Earnest Paul Ijjina and Krishna Mohan Chalavadi. 2017. Human action recognition in RGB-D videos using motion sequence information and deep learning. *Pattern Recognition* 72 (2017), 504–516.

[18] Mehrsan Javan Roshtkhari and Martin D Levine. 2013. Online dominant and anomalous behavior detection in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2611–2618.

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

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

[22] Federico Landi, Cees GM Snoek, and Rita Cucchiara. 2019. Anomaly locality in video surveillance. *arXiv preprint arXiv:1901.10364* (2019).

[23] Ce Li, Zhenjun Han, Qixiang Ye, and Jianbin Jiao. 2013. Visual abnormal behavior detection based on trajectory sparse reconstruction analysis. *Neurocomputing* 119 (2013), 94–100.

[24] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2013. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence* 36, 1 (2013), 18–32.

[25] Yuanyuan Li, Yiheng Cai, Jiaqi Liu, Shinan Lang, and Xinfeng Zhang. 2019. Spatio-temporal unity networking for video anomaly detection. *IEEE Access* 7 (2019), 172425–172432.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[27] Hong Liu, Juanhui Tu, and Mengyuan Liu. 2017. Two-stream 3d convolutional neural network for skeleton-based action recognition. *arXiv preprint arXiv:1705.08106* (2017).

[28] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6536–6545.

[29] Vijay Mahadevan, Wei-Xin LI, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly Detection in Crowded Scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1975–1981.

[30] Ramin Mehran, Alexis Oyama, and Mubarak Shah. 2009. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 935–942.

[31] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[33] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199* (2014).

[34] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6479–6488.

[35] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. 2019. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8797–8806.

[36] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[37] Di Zang, Zhenliang Chai, Junqi Zhang, Dongdong Zhang, and Jiujun Cheng. 2015. Vehicle license plate recognition using visual attention model and deep learning. *Journal of Electronic Imaging* 24, 3 (2015), 033001.

[38] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. 2020. A Comprehensive Study of Deep Video Action Recognition. *arXiv preprint arXiv:2012.06567* (2020).