Pattern Recognition Letters 145 (2021) 50–57

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Iterative weak/self-supervised classification framework for abnormal events detection

Bruno Degardin, Hugo Proença*

IT: Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal

ARTICLE INFO

Article history:
Received 13 June 2020
Revised 20 November 2020
Accepted 19 January 2021
Available online 4 February 2021

MSC:
41A05
41A10
65D05

Keywords:
Visual surveillance
Abnormal events detection
Weakly supervised learning

ABSTRACT

The detection of abnormal events in surveillance footage remains a challenge and has been the scope of various research works. Having observed that the state-of-the-art performance is still unsatisfactory, this paper provides a novel solution to the problem, with four-fold contributions: 1) upon the work of Sultani et al., we introduce one iterative learning framework composed of two experts working in the weak and self-supervised paradigms and providing additional amounts of learning data to each other, where the novel instances at each iteration are filtered by a Bayesian framework that supports the iterative data augmentation task; 2) we describe a novel term that is added to the baseline loss to spread the scores in the unit interval, which is crucial for the performance of the iterative framework; 3) we propose a *Random Forest* ensemble that fuses at the score level the top performing methods and reduces the EER values about 20% over the state-of-the-art; and 4) we announce the availability of the "UBI-Fights" dataset, fully annotated at the frame level, that can be freely used by the research community. The code, details of the experimental protocols and the dataset are publicly available at http://github.com/DegardinBruno/.

1. Introduction

The automatic detection of abnormal events in surveillance footage is still a tough challenge for various reasons. Not only the data acquisition environments are extremely diverse, with subjects imaged from varying distances, under multiple poses and partially occluded, but there is also a high complexity in resembling patterns among the different types of *abnormalities*.

Considering the difficulties in obtaining sufficient amounts of labelled *positive* data (i.e., with abnormal events), the majority of the existing approaches tackle the problem with the simple assumption of the existential of only *one-class*. This perspective, coined by Moya and Hush [12], requires modeling exclusively the ordinary dynamics present in the *normal* scenarios and considering any instance from distinct distributions (i.e., outliers) as an abnormality. In the generative model paradigm, [14] and [20] employed generative adversarial networks (GANs [5]) to learn generating solely normal events and assuming its incapability of generating non-analogous data since its reconstruction error will be higher than when normal events are used as input. Following this idea, [15] also applied GANs, but this time aiming to perceive *normal* pedestrian trajectories. Recently, [17] approached the problem from a weakly-supervision perspective: considering the exhaustive task of annotating at the frame level large video sequences, they proposed one method under the multiple instance learning paradigm, handling data annotated at the bag (video) level, achieving state-of-the-art performance.

This paper describes an iterative learning framework based on the weakly and self-supervised paradigms (Fig. 1), that consistently advances the state-of-the-art performance in abnormal events detection. We note that the proposed framework depends on one readily satisfiable condition: apart from having access to the weakly supervised set, an additional unlabelled set of videos should also be available (e.g., taken from *YouTube*). 1) Upon the work of [17], we use a weakly supervised dataset to infer a model that distinguishes between normal/abnormal segments (*weakly supervised* (WS) model). A novel term was added to the original loss function to enforce the spread of response scores in the unit interval. This term is the key that enables the discrimination between the most/less confidently classified instances; 2) we apply a Bayesian classifier responsible for filtering out the instances, coming from an unlabelled dataset, that are deemed to be the most reliably classified; 3) we consider that the high-belief instances are most likely to be correctly classified and employed these as the input of a second expert operating under the segment level (*self-supervised* (SS) network); 4) with the responses resulting from the SS, enable us to select the highly reliable videos that extend the learning set of the second iteration of our WS model.

* Corresponding author.
E-mail addresses: bruno.degardin@ubi.pt (B. Degardin), hugomcp@di.ubi.pt (H. Proença).

https://doi.org/10.1016/j.patrec.2021.01.031
0167-8655/

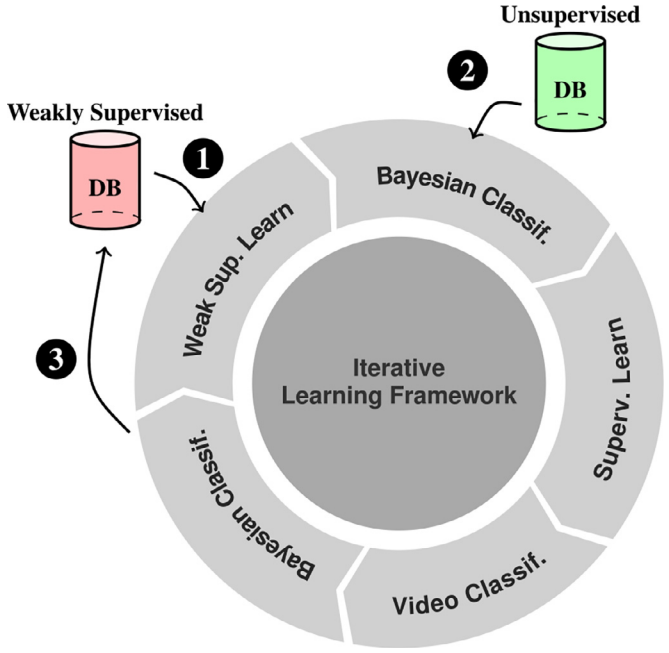

Fig. 1. Key insight the proposed method: we start from a weakly supervised classifier working at the video level (slightly changed with respect to [17]) that feeds a Bayesian classification module, responsible to select the most confidently correctly classified instances. These elements constitute the input of a self-supervised classifier (working at the block/segment level) that provides the instances used in the next iteration of the weakly supervised model.

The rationale is to repeatedly augment (based on unlabelled data) the learning sets used by the WS and SS experts. This way, the iterative process autonomously selects the novel instances to be added to the learning sets of both experts, being repeated while improvements in performance (in a validation set) are observed. A cohesive pictorial view of the whole framework is given in Fig. 2.

The remainder of this paper is divided into four sections as follows: Section 2 presents an overview of the most relevant works in the scope of abnormal event detection. Details and specifications of the proposed approach are provided in Section 3. In discuss in Section 4 the obtained results of our experiments and the conclusions are given in Section 5.

2. Related work

As in many other computer vision tasks that evolve spatiotemporal analysis, the use of three-dimensional convolutional networks has become state-of-the-art. Various works (e.g., [8,9,18]) report that 3D convolutions provide output volumes that simultaneously preserve the temporal and spatial information of the input. For video classification purposes, [9] trained several networks with large-scale datasets, reporting the advantages yielding from fusing both kinds of information at different stages of the processing pipeline. Tran et al. [18] proposed a video encoding scheme based on 3D ConvNets, coming out with a highly effective descriptor that has been used as the main source for several abnormal detection techniques (e.g., [4] and [17]).

For abnormal events detection purposes, [3] trained a spatial feature extractor followed by a temporal auto-encoder that is exclusively fed with *normal* events. Using the spatial features of every ten consecutive frames to feed the temporal encoder, they used nonlinear activation functions to minimize the reconstruction error in *normal* videos. The rationale is that, upon an *abnormal* scenario, the auto-encoder should not be capable to reconstruct the scene, at least in an effective way considering that it has never seen *abnormal* situations during the learning phase. Subsequently, following a similar strategy, [15] employed GANs to better train the discriminator expert. [6] considered improved trajectory features from the work of [19], that were fused to histograms of oriented gradients (HOG) and histograms of optical flows (HOF) features and used as input to an auto-encoder. Based on the concept of conditional GANs [7] (cGANs), [14] extracted the optical-flow from subsequent frames and trained two networks, both composed of a conditional generator and discriminator, with the purpose of generating frames from optical-flow and generating optical-flow from frames.

Considered the state-of-the-art, [17] proposed a method to learn anomalies through leveraging weakly labeled normal and abnormal videos under a multiple instance learning (MIL) paradigm. Dividing each video into non-overlapping temporal segments, each video represents a *bag*, where the temporal segments are used as instances. Since the specific information of a video containing an anomaly (positive bag) is not provided, the implemented loss function uses the maximum score in each bag to learn a ranking model, producing low scores and high scores for negative and positive instances, respectively, and improving the state-of-the-art performance significantly.

3. Proposed method

In a consistent manner, the notation adopted is as similar as possible to the used by Sultani et al. [17]. $\mathcal{V}_k^i$ denotes the ith segment of a bag (video) $\mathcal{B}_k$, where $k \in \{'a', 'n'\}$ refers respectively videos with *abnormal* and *normal* events. $\mathcal{B}^{(t)} = \{\mathcal{B}_1, \ldots, \mathcal{B}_{n_w}\}$ denotes a learning set of n_w videos at iteration t (used as learning data by the WS expert), and $\mathcal{V}^{(t)} = \{\mathcal{V}_1, \ldots, \mathcal{V}_{n_s}\}$ is the learning set composed of n_s segments, used by the SS expert.

3.1. Weakly-supervised network

The proposed WS model is based on [17]. Under the MIL paradigm [1], the idea is to consider the input videos as *bags*, annotated in a binary way: *positive* bags are videos that at some point have an abnormal segment, while *negative* bags assuredly contain only normal events. Considering that: 1) anomalies typically occur only for a short time; and 2) videos are sequences of segments, the scores should vary smoothly between consecutive segments, Sultani et al. proposed the following MIL ranking loss function:

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max\left(0, 1 - \max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right) + \lambda_1 \overbrace{\sum_i^{(n-1)} \left(f\left(\mathcal{V}_a^i\right) - f\left(\mathcal{V}_a^{i+1}\right)\right)^2}^{①} + \lambda_2 \overbrace{\sum_i^{n} f\left(\mathcal{V}_a^i\right)}^{②}, \quad (1)$$

where $f\left(\mathcal{V}_a^i\right)$ and $f\left(\mathcal{V}_n^i\right)$ denote the prediction scores of abnormal and normal segments, respectively, and max takes into account all video segments in each bag. In this formulation, with terms ① and ②, the temporal smoothness and sparsity are enforced, respectively.

Throughout our experiments, we immediately noticed that the loss function defined in (1) produced extremely peaked distributions either for *normal*/*abnormal* scores, turning hard to discriminate the most reliable correctly classified instances. Being heavily data-driven, these kinds of (neural-based) models are known to often produce output distributions peaked around the ground-truth values, which for classic classification tasks is even positively regarded. Second, it was also imperative to ensure that the scores for the most evidently *normal*/*abnormal* segments were the closest to the extremes of the unit interval, and spread as much as possible among intermediary values all the doubtful observations.

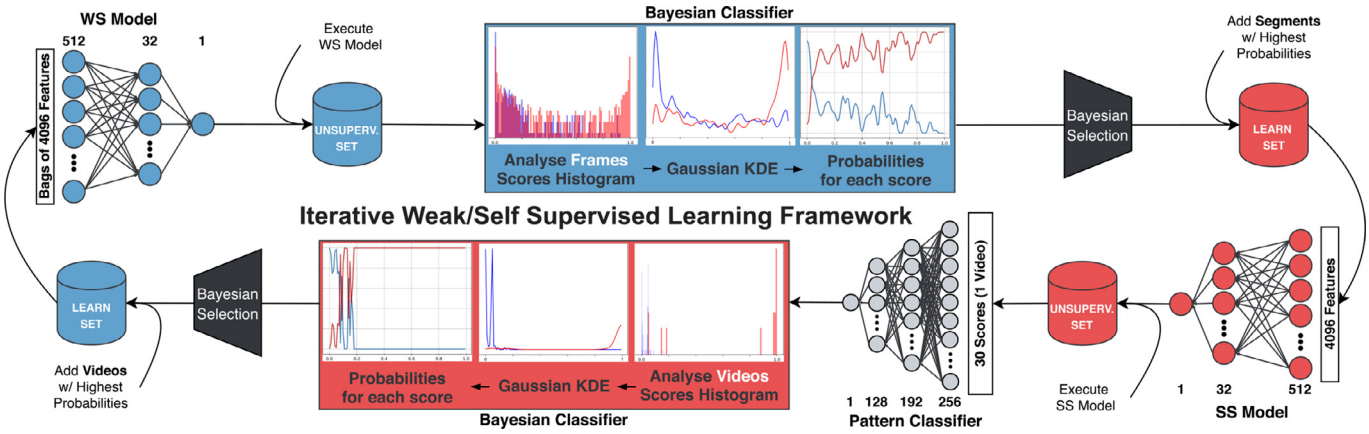

Fig. 2. Cohesive pictorial view of the iterative learning framework introduced in this paper. Two kinds of components are used, represented by the blue and red colors: firstly, [17]'s method was adapted in order to spread its output scores over the unit interval, which will facilitate the filtering by the Bayesian framework regarding the most reliably classified elements from the unlabelled data set and feed those to a self-supervised learning network operating under the segment level. Then, performing a cycle, the prediction scores by this expert are employed by a second Bayesian framework responsible for filtering the videos to be used in the following generation of the first framework. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

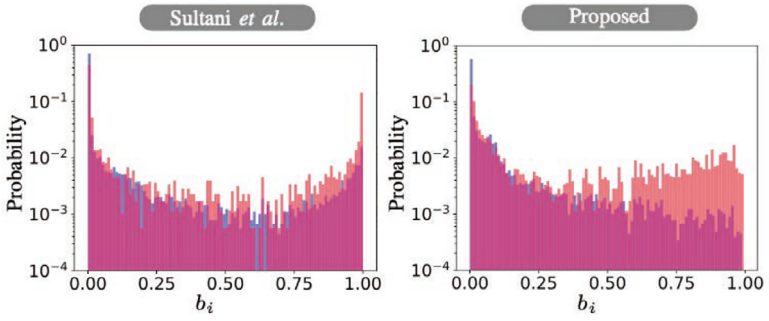

Fig. 3. Histograms comparison between the spread of $f(\mathcal{V}_u^i)$ over the unit interval, with respect to the applicability of the novel term in the loss function. Sultani et al.'s loss is illustrated in the left plot, while the right plot regards the values provided from the loss function proposed in this work. The vertical axis is plotted in log scale for visualization purposes.

Therefore, to achieve a better distribution score, an entropy-based term was added to the basis loss function, given by:

$$H(\mathcal{B}_u) = -\sum_{i=1}^{t} P\big(f(\mathcal{V}_u^i)\big)\ \log\Big(P\big(f(\mathcal{V}_u^i)\big)\Big), \tag{2}$$

where $\mathcal{B}_u \overset{\text{def}}{=} \{\mathcal{B}_a \cup \mathcal{B}_n\}$ describes all the videos in the batch, and the density of the $f(\mathcal{V}_u^i)$ scores are estimated by $P\big(f(\mathcal{V}_u^i)\big)$. The unit interval was divided into a fixed number of bins with equal width, counting the number of values in each bin:

$$b_i = \sum_{j=1}^{n} \mathbb{1}_{\{\frac{i-1}{n} \le f(\mathcal{V}_u^j) \le \frac{i}{n}\}}, \forall i \in \{1, \ldots, t\}, \tag{3}$$

where $\mathbb{1}_{\cdot}$ represents the indicator function. Values were normalized to have its summation equal to one, after obtaining the b_i statistic, i.e., estimating a distribution probability: $b_i^* = \frac{b_i}{\sum_j b_j}$. In practical terms, $P\big(f(\mathcal{V}_u^i)\big)$ was approximated by making use of the b_i^* values. The major impact of the novel entropy term is demonstrated in Fig. 3, with the vertical axis denoting the b_i^* values per bin and the horizontal axis providing the codomain of the classifier.

Incorporating (2) to the loss function (1) and also taking into account the regularization terms that minimizes the model weights, our final objective function becomes:

$$\mathcal{L}(\mathcal{W}) = l(\mathcal{B}_a, \mathcal{B}_n) - \lambda_3 \overbrace{H(\mathcal{B}_u))}^{③} + \lambda_4 \|\mathcal{W}\|_F, \tag{4}$$

where $\mathcal{W}$ denotes the model weights and ③ avoids peaked density values near the extremes of the unit interval for both *normal*/*abnormal* events.

As noted earlier, the rationale of (4) is that more dispersed and less peaked distributions of prediction scores will result in higher entropy values. Despite being less frequent to obtain scores close to 0 and 1, the crucial issue is to assure that only the most undoubtedly *normal*/*abnormal* instances reach those values close to the extremes, being the key feature in the selection of instances, which are iteratively inserted into the next generation of learning sets, as a result of the Bayesian framework.

3.2. Self-supervised network

Following a standard classification paradigm, the SS network operates under the segment level, which receives C3D [18] feature vectors of 4,096 components representing segments of a video (16 frames/each), predicting the *abnormality*, i.e., the likelihood of a segment containing an *abnormal* event. As demonstrated in Fig. 4, with a 3-layer fully connected architecture, this expert is composed of 512 units in the first layer, 32 units in the second one, and 1 unit as the output layer. Employing ReLU [13] as the acti-

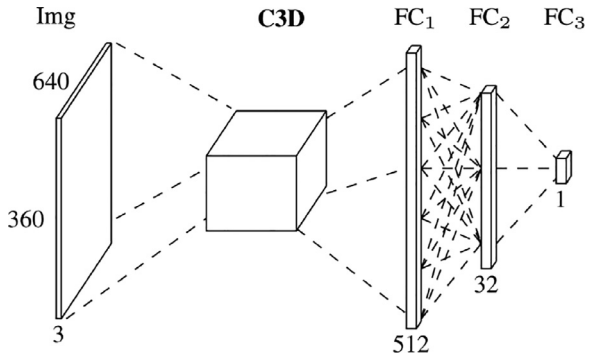

Fig. 4. Structure of the self-supervised (SS) network that discriminates between *normal*/*abnormal* segments. The C3D [18] descriptor is used as feature encoder, that feeds a three layer fully connected model.

vation function for the first two layers and a sigmoid activation in the output neuron, the learning process of the model is driven by a binary cross-entropy loss.

3.3. Bayesian classifiers

For each classified instance, among the unsupervised data, the Bayesian classifiers allow obtaining its degree of belief. The foundation of this approach is that only the extremely high believed instances should be included in the next generation of the learning set, in a self-supervised fashion for either the WS and SS networks. Two Bayesian classifiers are embedded in our method's framework in order to achieve self-supervision: 1) the *weak* one (P_w) responsible for filtering out the segments that should be used in the SS training, based upon the received scores produced by the WS expert; and 2) the *strong* counterpart (P_s) in charge of selecting the videos that should be used in the WS expert's next-generation, accordingly to the received scores produced by the SS model.

$$P_w\big(y|f(\mathcal{V}^i)\big) = \frac{P\big(f(\mathcal{V}^i)|y\big)\ P(y)}{P\big(f(\mathcal{V}^i)\big)} \tag{5}$$

where $y \in \{'a', 'n'\}$ represent the *abnormal*/*normal* classes. We applied a Gaussian kernel density estimator, with Scott's rule [16] for bandwidth selection, with the aim of approximate more accurately the conditional densities $P\big(f(\mathcal{V}^i)|y\big)$. Due to the existence of a significant imbalance regarding the number of *abnormal* instances compared to *normal* ones, the priors were empirically adjusted to $P('a') = P('n') = 0.5$.

Fig. 5 illustrates the comparison of the posteriors per class, according to the histograms in Fig. 3, which regards the Sultani et al.'s loss and ours when the entropy term was also considered. The *abnormality* scores are represented by the red lines, and the normal ones by the blue lines. The immediate impact of our approach is to attain a more evident dispersion near the extremes of the unit interval between the degrees of belief for both classes, particularly in the upper extremity.

Formally, the next generation of the SS learning set will include the *i*th segment according to the rule:

$$\mathcal{V}^{(t+1)} \stackrel{\text{def}}{=} \{\mathcal{V}^i \iff P_w\big(y|f(\mathcal{V}^i)\big) \geq \tau_1\}, y \in \{'a', 'n'\}, \tag{6}$$

i.e., given a segment's prediction score, it is selected if its posterior for either the 'n'/'a' classes is higher than a threshold.

Regarding the self-supervision of the WS expert (step 3 in Fig. 1), the *strong* Bayesian classifier receives the video segments scores and should decide at the video (bag) level, globally classifying the video as *normal*/*abnormal*:

$$P_s\big(y|f(\mathcal{B}^i)\big) = \frac{P\big(f(\mathcal{B}^i)|y\big)\ P(y)}{P\big(f(\mathcal{B}^i)\big)} \tag{7}$$

where the *abnormal*/*normal* classes are denoted as $y \in \{'a', 'n'\}$. The poor levels of performance of simple fusion rules (such as *max* or *prod*) in obtaining the degree of belief per video, led us to the application of a simple feed-forward classification module with a 4-layer fully connected architecture designed with 256 : 196 : 128 : 1 for estimating $P_s\big(y|f(\mathcal{B}^i)\big)$. This network is trained with the SS expert's result scores from the validation set in each generation to be able to adapt itself according to the SS model's evolution. Finally, similarly to the *weak* Bayesian classifier, only the videos that present the extreme degrees of belief are inserted into the next generation of the learning set:

$$\mathcal{B}^{(t+1)} \stackrel{\text{def}}{=} \{\mathcal{B}^i \iff P_s\big(y|f(\mathcal{B}^i)\big) \geq \tau_2\}, \forall i \in \mathcal{B}^{(t)}, y \in \{'a', 'n'\}, \tag{8}$$

In accordance with the performance observed in a validation set, the thresholds τ_i were adjusted empirically.

4. Experiments and discussion

4.1. Datasets, baselines and empirical evaluation protocol

Our experiments were conducted in three datasets: 1) specifically concerned about the detection of *fighting* events in surveillance footage, we mined 1,000 videos (collected from *Youtube* and *LiveLeak*), where 216 contain a real-life *fighting* scenario, and the remaining 784 contain only normal events. This dataset was manually annotated at the frame level and is freely available. It contains video data from both outdoor/indoor environments, with subjects appearing under completely uncontrolled poses, lighting conditions, at very different scales, and often under severe occlusions. 2) the UCF-Crime [17], which contains 1,900 untrimmed surveillance videos, including 13 types of abnormalities and *normal* scenarios, providing video level annotation (learning) and at the segment level (test). It is currently the most extensive publicly available data source for abnormal events detection. It was collected from the web and contains highly heterogeneous data, collected at different light spectra (near-infrared and visible), varying scale, perspectives, distances, and in indoor/outdoor environments. 3) the UCSD dataset [11], acquired by a stationary camera mounted at an elevation, overlooking pedestrian walkways. In this set, the crowd density in the walkways varies, ranging from sparse to very crowded. *Normal* video contains only pedestrians, while *abnormal* events are due to either the circulation of non-pedestrian entities or due to anomalous pedestrian motion patterns. This one is the less heterogeneous of the three sets tested, with all video files being collected by the same device, using near-infrared light, and subjects appearing at similar (high) scales and perspectives.

Each video was resized to 360 × 640 pixels, with a normalized frame rate of 30 fps, and trimmed into *sub-videos* with a constant length of 16 s, and corresponding annotations manually adjusted. The *UBI-Fights* set was split into three disjoint subsets: 80% (for learning), 5% (for validation purposes) and 15% (for the test phase), while for the *UCF-Crime* and *UCSD* sets we used the default learning sets, with 50% of the test set (randomly chosen) for validation purposes. Next, for every 16 frame-clip (i.e., one segment) for the three datasets, we employed the sixth fully-connected layer (FC6) of the C3D network [18] in order to extract the spatiotemporal descriptors. Regarding the parameterization of our solution, we empirically set $\tau_1 = \tau_2 = 0.99$ (*Fights*) and 0.999 (*UCF-Crime* and *UCSD*), $\lambda_1 = \lambda_2 = 8 \times 10^{-5}$, $\lambda_3 \in [0, \frac{\log(ep+1)}{30}]$, regarding the training epoch ep, and $\lambda_4 = 1$.

Concerning the baselines, five methods, regarding both one-class and binary classification, were considered to represent the state-of-the-art: [3,6,14,17,20], all described in Section 2. Apart from these, a quadratic kernel SVM operating under the single-class paradigm was also tested, which was fed by the same C3D

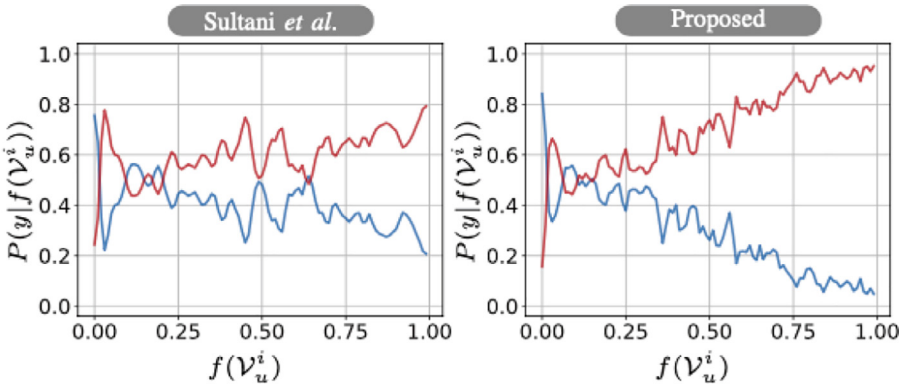

Fig. 5. Effect of the novel term added to the [17] loss function and compared between the posteriors $P\left(y|f(\mathcal{V}^i)\right)$ estimated by the Bayesian classifier with the application of the Sultani et al. original formulation (at left) and ours (at right).

feature vectors as our WS and Sultani et al. models. Three performance measures are reported: the decidability index (d'), the area under the curve (AUC), and the equal error rate (EER).

In order to ensure a fair evaluation between the state-of-the-art and our proposal, taking into consideration that one premise of our method is the availability of unsupervised data, 50% of the ground-truth labels were disregarded from the learning sets and considered those subsets as *unlabelled*. Therefore, we also removed these instances from the learning sets of the other approaches, with the purpose of using rigorously the same amount of labeled data for the learning phase.

All our experiments were conducted according to a bootstrapping-like strategy that is widely adopted in computer vision/pattern recognition experiments (e.g. [10]). Having n instances available, the bootstrap randomly selects (without replacement in our case) $0.9n$ of them, creating a sample composed by 90% of the available data. This sample is further disjointly divided into two subsets: one for learning purposes and the other for performance evaluation. The bootstrapping-like draw was repeated 10 times per data set, creating 10 subsets of each one. Next, the experiments (model learning and performance evaluation) were conducted in each subset, which enables to obtain the average and standard deviation performance values at all operating points.

4.2. Results and discussion

4.2.1. State-of-the-art comparison

The immediate observation was the extremely poor performance that single-class techniques achieve in this type of problem, in conformity with the conclusion drew by Sultani et al.. These techniques fail in an almost catastrophic way to appropriately establish a boundary between normality and abnormality in test data, particularly in cases of outdoor environments. From our observation, the heterogeneity of the expected inputs in those datasets is extremely high, not forgetting that these frameworks ultimately require tremendous additional amounts of training data to become minimally effective.

Concerning our approach, Fig. 6 provides the learning curve of our solution, i.e., the evolution in WS/SS models performance, with respect to the iteration. For the *UBI-Fights* and *UCF-Crime* datasets, the SS expert achieves its maximum AUC value at the end of the third and second iteration, respectively, while the WS model obtained its optimal point at the forth iteration with a clear margin in the *UBI-Fights* set, and in the *UCF-Crime* only by a residual difference. With regard to the results in the latter set, the WS expert performance is justified by the considerably more difficult task of extracting simultaneously consistent patterns among 13 different types of abnormalities, attaining practically similar results as the Sultani et al. method in all iterations.

The *UCSD* set provided the best performance values, both for our proposal and the Sultani et al. baseline. In this case, the first generation of the WS and SS models attained results clearly worse than the baseline performance, which was only surpassed by the subsequent generations of both models, when working with additional amounts of data. Overall, this dataset can be considered as relatively easier than the remaining ones, as a result of containing data of a single scene, acquired by a stationary device, and in relatively homogenous lighting conditions. Here, the observed failure cases were in most circumstances due to some ill-definition of an *abnormal* event, for some *anomalous pedestrian trajectories* poorly represented in the learning set.

Fig. 7 compares the ROC curves observed for all methods, in the *UBI-Fights* (left plot), *UCF-Crime* (central plot) and *UCSD* (right plot) sets. Overall, our solution achieved clearly better results than any competitor, with the exception of a small region of the performance space in the *UBI-Fights* dataset, where Sultani et al.'s method got results to ours. In an interesting way, the performance improvements were more obvious in the *UCF-Crime* set than in the *UBI-Fights*, despite that our method was developed taking into account the specific detection of fighting scenarios. In all cases, Sultani et al. was the runner up proposal, which corresponds to the results previously reported. Nevertheless, the noticed gap between the reported results in [17] and here enables the conclusion concerning the extremely high sensitivity of this approach with respect to the amounts of learning data used. The remaining approaches were considered to fail in a disastrous way of detecting abnormal events, at least with these amounts of data for learning purposes, and with a extremely environmental heterogeneity as considered in these experiments.

The application of single-class learning techniques appears to be appropriate exclusively when providing learning data that features highly similar lighting conditions and camera pose in the learning and test set, which happens in the *UCSD* dataset. In this case, even though our SS proposal got again the best performance among all techniques, [17] and [3] techniques attained similar effectiveness in some regions of the performance space.

Regarding the solution described in this paper, the experiments support its effectiveness, in particular when the initial data set (at iteration 0) represent the major data variation factors that the system should handle. In cases where that set is not representative of the typical data variation factors (i.e., lack of specific lighting conditions, or particular perspectives/scales), obvious decreases in performance were observed (with AUC values reducing over 5%). In our view, this is due to the lack of generalisation capabilities of

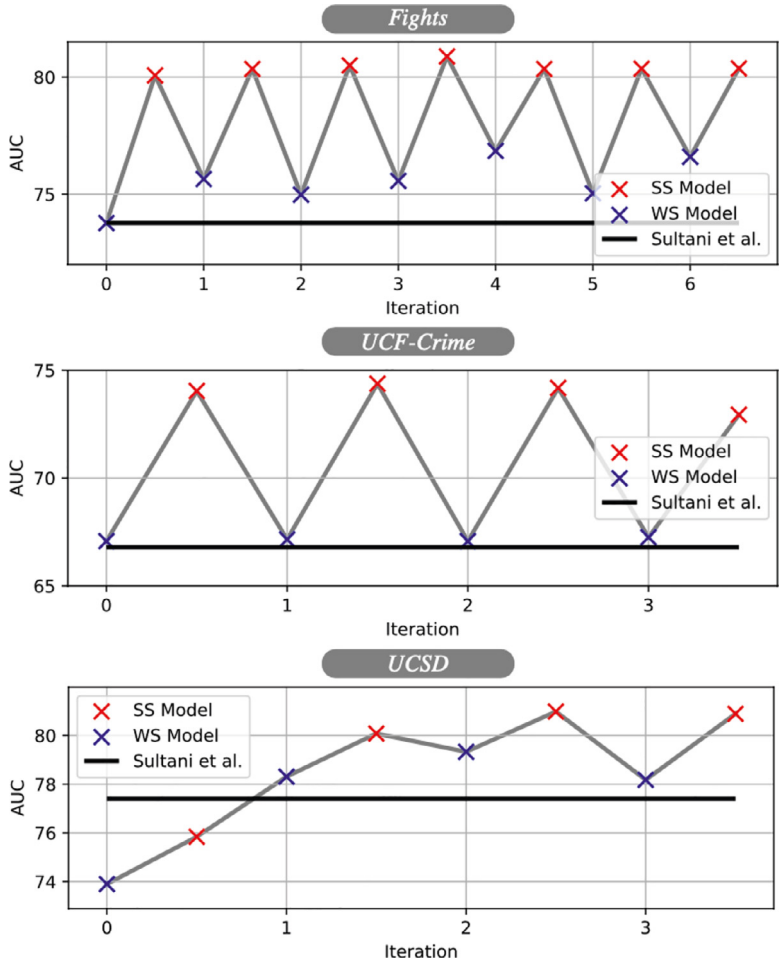

Fig. 6. Performance improvements in the validation set of our frameworks regarding its iteration, observed in each dataset. The SS expert is denoted with the red points and the WS expert with the blue points. As our main baseline, Sultani et al.'s performance is represented with the continuous black lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

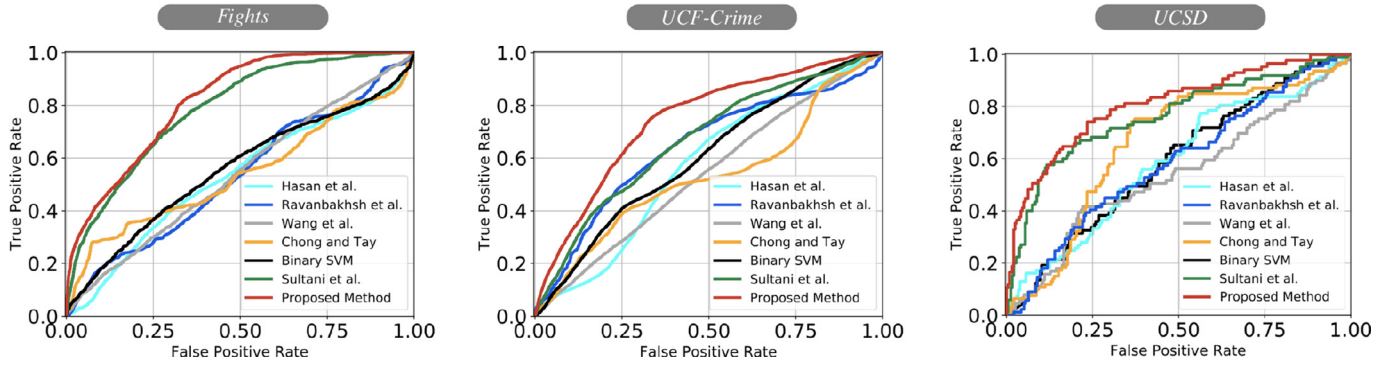

Fig. 7. ROC curves comparison obtained by the seven approaches considered in our experiments, concerning the *UBI-Fights* (left plot), *UCF-Crime* (center plot) and *UCSD* (right plot) datasets.

the Bayesian classifiers that then face difficulties in classifying the data with novel data variation factors as "*normal/abnormal*".

4.2.2. Ensemble classifier

Upon our state-of-the-art comparison, we noticed that the tested methods have very different typical failure cases, which suggests that a *meta-classifier* ensemble would possibly maximize the improvements due to fusion. Hence, we measured the linear correlation between the responses produced by the eight methods for the *UCF-Crime* and *Fights* datasets, which were considered the most appropriate for our purposes due to the heterogeneity of the data they contain. Results are shown in the left plot of Fig. 8, with an 8 $\times$ 8 matrix describing the Pearson correlation coefficients between our WS and SS models, Sultani et al. (S), Hasan et al. (H), Ravanbakhsh et al. (R), Wang et al. (W), Chong and Tay (C) and the SVM classifier (V). The upper diagonal provides the correlation coefficients for the *UCF-Crime* set, while the corresponding *Fights* results are provided in the lower diagonal of the matrix. The levels of correlation in both sets were approximately the same, even though slightly higher values were observed for the *Fights* set.

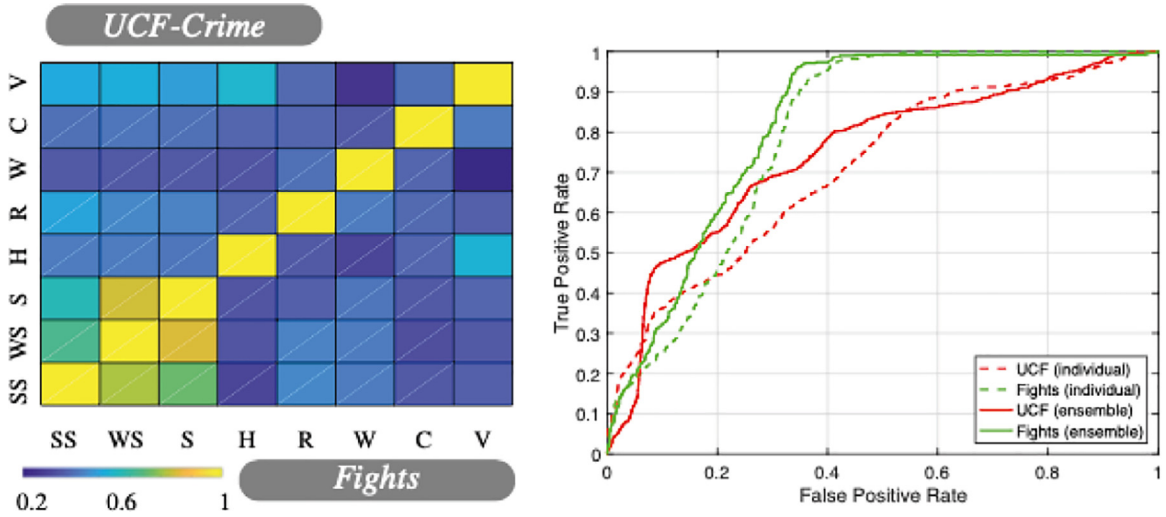

Fig. 8. At left: linear correlation (Pearson coefficient) among the predictions produced by the proposals analyzed. At right: performance improvements for the *UBI-Fights* and *UCF-Crime* sets when employing a Random Tree ensemble, which fuses at the score level the output of the WS/SS/[17] experts (continuous lines) concerning the best individual expert (dashed lines).

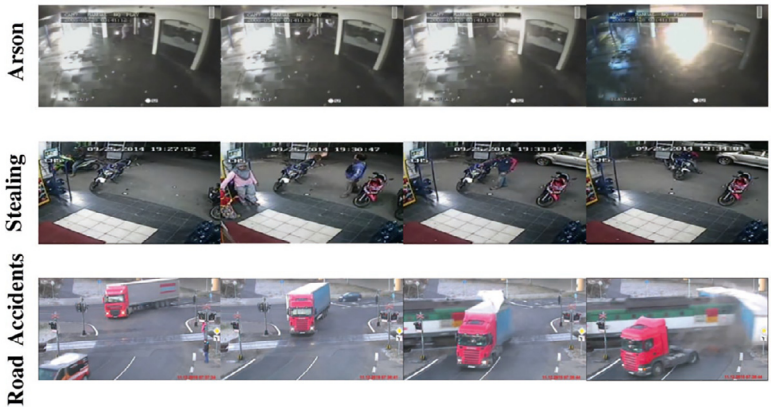

Fig. 9. Qualitative analysis of our results.Top rows: examples of the *Arson* and *Stealing* classes, where the most significant improvements of our solution with respect to Sultani et al. were observed. Bottom row: example of the *Road Accidents* case, where no relevant improvements were attained.

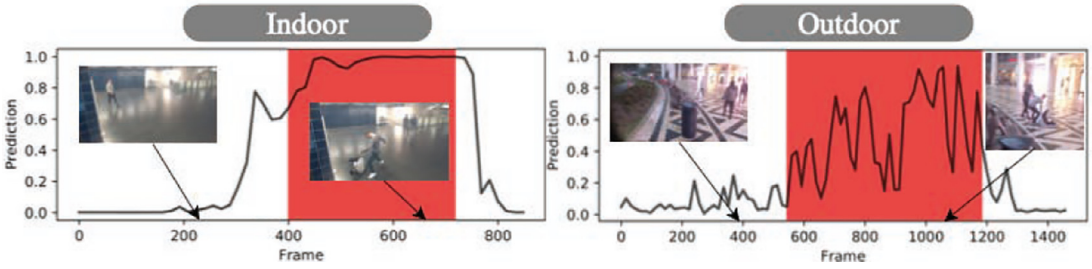

Fig. 10. Gap in performance typically observed between indoor and outdoor environments. The dynamics of outdoor conditions augment considerably the challenges in obtaining smooth responses.

The low levels of linear correlation motivated us to exploit the improvements in performance due to fusion. We used a Random Forest model [2] to fuse at the score level the responses provided by the best-performing algorithms (i.e., our SS/WS models and [17]). The results are provided at the right plot in Fig. 8, and turn evident a slight improvement in effectiveness with respect to the best individual expert, attaining AUC values of about 0.819 (*Fights*) and 0.769 (*UCF-Crime*). The final results are summarized in Table 1, providing the AUC, decidability (d'), and EER values for all the methods tested individually and also for the ensemble.

In terms of a qualitative comparison between our solution and the baselines, we concentrated our analysis in the *UCF-Crime* data set and in the Sultani et al.solution. In this case, improvements were maximally evident for the *Arson* and *Stealing* cases, that are illustrated in Fig. 9. In opposition, no significant improvements of our solution with respect to the baseline were observed for the *Road Accidents* class.

As a concluding remark, even though the proposed method (and the ensemble) consistently advance the state-of-the-art, there is still an evident gap between the performance attained in indoor scenarios (where lighting conditions are controlled) and outdoor

Table 1
Performance summary of the proposed method with respect to the state-of-the-art. The final rows in the *Fights* and *UCF-Crime* datasets provide the performance of an ensemble Random Forest that fuses at the score level the responses of our SS/WS models to [17].

Method	AUC	d'	EER
UBI-Fights			
[6]	0.528 ± 0.003	0.194 ± 0.007	0.466 ± 0.002
[14]	0.533 ± 0.003	0.147 ± 0.001	0.484 ± 0.002
[20]	0.540 ± 0.002	0.164± 0.008	0.475 ± 0.002
[3]	0.541± 0.003	0.059 ± 0.001	0.480 ± 0.001
Binary SVM Classifier	0.556 ± 0.003	0.128 ± 0.012	0.443 ± 0.003
[17]	0.787 ± 0.002	0.738 ± 0.011	0.294 ± 0.002
Proposed Method	**0.819** ± 0.001	**0.986** ± 0.008	**0.284** ± 0.001
Ensemble	**0.846** ± 0.001	**1.108** ± 0.008	**0.216** ± 0.001
UCF-Crime			
[6]	0.573 ± 0.001	0.167 ± 0.006	0.424 ± 0.001
[14]	0.642 ± 0.001	0.423 ± 0.005	0.376 ± 0.001
[20]	0.539 ± 0.001	0.104 ± 0.004	0.472 ± 0.001
[3]	0.532 ± 0.002	0.138 ± 0.006	0.484 ± 0.001
Binary SVM Classifier	0.604 ± 0.001	0.377 ± 0.003	0.441 ± 0.001
[17]	0.668 ± 0.001	0.523 ± 0.004	0.375 ± 0.001
Proposed Method	**0.744** ± 0.001	**0.764** ± 0.005	**0.305** ± 0.001
Ensemble	**0.769** ± $9e^{-4}$	**0.890** ± 0.005	**0.266** ± $9e^{-4}$
UCSD			
[6]	0.591 ± 0.011	0.309 ± 0.040	0.425 ± 0.014
[14]	0.582 ± 0.011	0.139 ± 0.019	0.443 ± 0.006
[20]	0.545 ± 0.011	0.113 ± 0.039	0.473 ± 0.009
[3]	0.660 ± 0.010	0.264 ± 0.030	0.343 ± 0.009
Binary SVM Classifier	0.590 ± 0.012	0.310 ± 0.043	0.441 ± 0.008
[17]	0.758 ± 0.009	0.811 ± 0.059	0.302 ± 0.012
Proposed Method	**0.809** ± 0.010	**1.112** ± 0.060	**0.252** ± 0.011

environments, where the scores typically are much more irregular than in the indoor setting. This problem is illustrated in Fig. 10 and should motivate further developments in this technology.

5. Conclusions and further work

In this paper, we described a solution for abnormal events detection in surveillance footage that fuses the weakly and self-supervised learning paradigms. We introduced an iterative learning framework composed of two experts (based upon the weakly and self-supervised paradigms) feeding data to each other, where two Bayesian frameworks infer the most likely correctly classified instances from each expert. Such frameworks filter the input instances of the weakly and self-supervised experts and iteratively provide additional data to each other. More importantly, having observed that the proposed experts provide scores that are poorly correlated to the values generated by Sultani et al. [17], we used a Random Forest model that fuses at the score level the outputs of the top-three best performing models. The final ensemble outperforms the state-of-the-art in both datasets, reducing the EER from 15% (*UCF-Crime*) to 24% (*UBI-Fights*). As an additional contribution, the *UBI-Fights* dataset is fully annotated at the frame level and publicly available[1] for the research community. The source code of the method and the datasets are freely available and can be used to support further developments in the challenging problem of abnormal events detection.

As possible directions for further work, one direction will be the analysis of the properties of the initial generation of the data set, to maximize effectiveness and improvements in subsequent iterations. Also, other options could be used as a replacement of the Bayesian classifiers, such as the creation and mapping of latent codes into low-dimensional manifolds, from where simple rules based in Euclidean distance could be used to select the most confidently classified *normal*/*abnormal* instances.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the FCT/MEC through National Funds and Co-Funded by the FEDER-PT2020 Partnership Agreement under Project UIDB/50008/2020, Project POCI-01-0247-FEDER-033395 and in part by the C4: Cloud Computing Competence Centre.

References

[1] J. Amores, Multiple instance classification: Review, taxonomy and comparative study, Artif. Intell. 201 (2013) 81–105.
[2] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[3] Y.S. Chong, Y.H. Tay, Abnormal event detection in videos using spatiotemporal autoencoder, International Symposium on Neural Networks 189–196, Springer, 2017.
[4] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 1933–1941, 2016.
[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[6] M. Hasan, J. Choi, J. Neumann, A.K. Roy-Chowdhury, L.S. Davis, Learning temporal regularity in video sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 733–742.
[7] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[8] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach.Intell. 35 (2012) 221–231.
[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[10] K. K. Hollingsworth, K. Bowyer, P. Flynn, Improved iris recognition through fusion of hamming distance and fragile bit distance, IEEE Trans. Pattern Anal. Mach.Intell. 33 (2011) 2465–2476.
[11] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern Anal. Mach.Intell. 36 (2014) 18–32.
[12] M. Moya, D. Hush, Network constraints and multi- objective optimization for one-class classification, Neural Netw. 9 (1996) 463–474.
[13] V. Nair, G. Hinton, Rectified linear units improve restricted boltzmann machines Vinod Nair, in: Proceedings of the IEEE International Conference on Machine Learning, 2010, pp. 807–814.
[14] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, N. Sebe, Abnormal event detection in videos using generative adversarial nets, IEEE, 2017. 2017 IEEE International Conference on Image Processing (ICIP) 1577–1581
[15] P. Roy, G.A. Bilodeau, Adversarially learned abnormal trajectory classifier, in: 2019 16th Conference on Computer and Robot Vision (CRV) 65–72, IEEE, 2019.
[16] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons, 2015.
[17] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479–6488.
[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[19] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.
[20] T. Wang, M. Qiao, Z. Lin, C. Li, H. Snoussi, Z. Liu, C. Choi, Generative neural networks for anomaly detection in crowded scenes, IEEE Trans. Inf. ForensicsSecur. 14 (2018) 1390–1399.

[1] http://socia-lab.di.ubi.pt/EventDetection