Human Activity Analysis: Iterative Weak/Self-Supervised Learning Frameworks for Detecting Abnormal Events

Bruno Degardin and Hugo Proença
IT - Instituto de Telecomunicações
University of Beira Interior, Portugal
bruno.degardin@ubi.pt, hugomcp@di.ubi.pt

Abstract

Having observed the unsatisfactory state-of-the-art performance in detecting abnormal events, this paper describes an iterative self-supervised learning method for such purpose. The proposed solution is composed of two experts that - at each step - find the most confidently classified instances to augment the amount of data available for the next iteration. Our contributions are four-fold: 1) we describe the iterative learning framework composed of experts working in the weak/self-supervised paradigms and providing learning data to each other, with the novel instances being filtered by a Bayesian framework; 2) upon Sultani et al. *[14]'s work, we suggest a novel term the loss function that spreads the scores in the unit interval and is important for the performance of the iterative framework; 3) we propose a late decision fusion scheme, in which an ensemble of* Decision Trees *learned from bootstrap samples fuses the scores of the top-3 methods, reducing the EER values about 20% over the state-of-the-art; and 4) we announce the "*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 analysis of human activities is still a tough challenge. In particular, detecting *abnormal* events in urban environments is challenging due to various reasons: not only the data acquisition environments are extremely diverse, but resembling patterns among the various possible types of *abnormalities* is particularly difficult, due to their heterogeneity.

Considering the difficulties in obtaining labelled *positive* instances (abnormal events), most of the existing methods approach the problem from the *one-class* classification perspective (term coined by Moya and Hush [9]), modelling exclusively normal events and assuming that elements from different distributions (i.e., outliers) would represent *abnormalities*. Among other examples, [11] and [17] used generative adversarial networks (GANs [4]) to generate normal data, assuming that non-analogous events fed to the generator will provide higher reconstruction losses than normal sequences. Also, Roy and Bilodeau [12] used GANs to perceive *normal* pedestrian trajectories. More recently, Sultani *et al.* [14] proposed one method based in the multiple instance learning (MIL) paradigm and using data annotated at the bag (video) level, coming out with a model that is considered the state-of-the-art.

In this paper we describe one iterative learning framework based in the weakly and self-supervised paradigms, composed of two experts feeding data to each other. 1) Upon Sultani *et al.* [14]'s method, we infer a model that distinguishes between normal/abnormal segments (*weakly supervised* (WS) model). Importantly, a novel term was added to the original loss function to spread of response scores in the unit interval, which enables the discrimination between the most/less confidently classified instances; 2) we use a Bayesian classifier to filter out the less reliably classified instances in an unlabelled dataset; 3) we deem that the high-belief instances are correctly classified and used these as input of a second network working at the segment level (*self-supervised* (SS) expert); 4) based in the responses produced by the SS model, we select the instances that will extend the WS learning set in the second iteration.

The rationale is to iteratively augment (based on unlabelled data) the learning sets used by the WS and SS experts. The iterative process selects the novel instances added to the learning sets of both experts, and is repeated while improvements in validation performance are observed. A pictorial view of the whole framework is given in Fig. 1.

978-1-7281-9186-7/20/$31.00 ©2020 IEEE

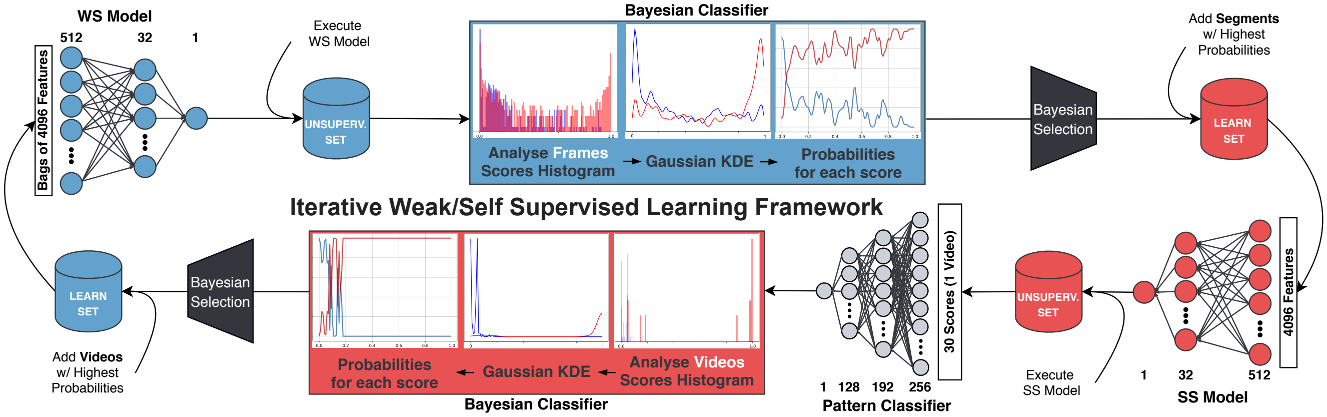

Figure 1: Cohesive view of the iterative learning framework proposed in this paper. The blue and red colors represent the two kinds of components used: at first, we adapted Sultani *et al.* [14]'s method to spread its output scores in the unit interval. Then, a Bayesian framework filters the instances of an unlabelled data set, and feeds the most reliably classified elements to a self-supervised learning network working at the segment level. Again, the responses predicted by this expert are used by a second Bayesian framework that filters the videos to be used in the next iteration of the first framework.

The remainder of this paper is organized as follows: Section 2 summarizes the most relevant works in the scope of our work. Section 3 provides the details of the proposed method. In Section 4 we discuss the obtained results and the conclusions are given in Section 5.

2. Related Work

Various works (e.g., [7], [8] and [15]) have reported that 3D convolutions provide output volumes that simultaneously preserve the spatiotemporal input features. Karpathy *et al.* [8] also reported the advantages yielding from fusing both kinds of information at different stages of the processing pipeline. Tran *et al.* [15] proposed a video encoding scheme based on 3D ConvNets, coming out with a descriptor that has been used as main source for various abnormal detection techniques (e.g., [3] and [14]).

Chong and Tay [2] trained a spatial feature extractor followed by a temporal auto-encoder that is fed with *normal* events. The rationale is that, upon an abnormal event, the auto-encoder should not be able to reconstruct it, at least in an effective way as for the *normal* events seen during the learning phase. Roy and Bilodeau [12] followed a similar strategy, based in GANs to better train the discriminator expert. Hasan *et al.* [5] considered improved trajectory features from Wang and Schmid [16]'s work, fused to histograms of oriented gradients (HOG) and histograms of optical flows (HOF) features to feed an auto-encoder. Based in conditional GANs [6] (cGANs), Ravanbakhsh *et al.* [11] extracted the optical-flow from consecutive frames and trained two networks, one to generate optical-flow from frames and another to generate frames from optical-flow, both composed of a conditional generator and discriminator.

Sultani *et al.* [14] proposed the method that is considered the state-of-the-art. They exploited both normal and abnormal videos, learning anomalies in a multiple instance learning (MIL) paradigm. Each video was considered a *bag*, and divided into non-overlapping temporal segments, used as instances. As the specific information of the positive bag (video containing an anomaly) is not known, the implemented ranking loss uses the maximum scores of each video to produce a ranking model, obtaining low scores and high scores for negative and positive instances, significantly advancing the state-of-the-art performance.

3. Proposed Method

For consistency purposes, we adopt a notation as close as possible to Sultani *et al.* [14]'s. $\mathcal{V}_k^i$ denotes the ith segment of a video (bag) $\mathcal{B}_k$, where $k \in \{'a','n'\}$ refer videos with *abnormal* and *normal* events. $\boldsymbol{\mathcal{B}}^{(t)} = \{\mathcal{B}_1, \dots, \mathcal{B}_{n_w}\}$ is a learning set of n_w videos at iteration t (used as learning data by the WS model). Similarly, $\boldsymbol{\mathcal{V}}^{(t)} = \{\mathcal{V}_1, \dots, \mathcal{V}_{n_s}\}$ is the learning set used by the SS model.

3.1. Weakly-supervised Network

The proposed WS model is based in the work of Sultani *et al.* [14]. Under the MIL paradigm, we consider the videos as *bags*, annotated in a binary way: *positive* instances have an abnormal event at some point, while *negative* instances assuredly contain only normal segments. Sultani *et al.* proposed the following MIL ranking loss function:

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

where $f\left(\mathcal{V}_a^i\right)$ and $f\left(\mathcal{V}_n^i\right)$ represent the prediction scores of abnormal and normal segments and $\max$ considers the video segments in each bag. In this formulation, $\textcircled{1}$ enforces the temporal smoothness, while $\textcircled{2}$ is the sparsity term.

During our experiments, we noticed that the loss function described in (1) typically produces extremely peaked distributions for both the *normal/abnormal* scores, turning hard to perceive the most confidently correctly classified instances. For our purposes, it was also important to assure that the responses for the most evidently *normal/abnormal* segments were close to the extremes of the unit interval, while all the scores for uncertain observations should spread as much as possible among intermediary values. Hence, an entropy-based term was added to the 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 $\mathcal{B}_u \overset{\text{def}}{=} \{\mathcal{B}_a \cup \mathcal{B}_n\}$ denotes all the videos in the batch, and $P\big(f(\mathcal{V}_u^i)\big)$ approximates the density of the $f(\mathcal{V}_u^i)$ scores. We divided the unit interval into bins of equal width and counted the number of values in each bin:

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

where $\mathbb{1}_{.}$ represents the characteristic function. After obtaining the b_i statistic, values were normalized to have sum equal to one, i.e., approximating a distribution probability: $b_i^*=\frac{b_i}{\sum_j b_j}$. In practice, the b_i^* values were used to approximate $P\big(f(\mathcal{V}_u^i)\big)$. The primary effect of the novel term is illustrated in Fig. 2, with the horizontal axis denoting the codomain of the classifier and the vertical axis provides the b_i^* values per bin.

Fusing (1)-(2) and considering also the regularization terms that minimizes the model weights, our complete objective function becomes:

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

where $\mathcal{W}$ represents the model weights and $\textcircled{3}$ avoids peaked distributions for *normal/abnormal* events.

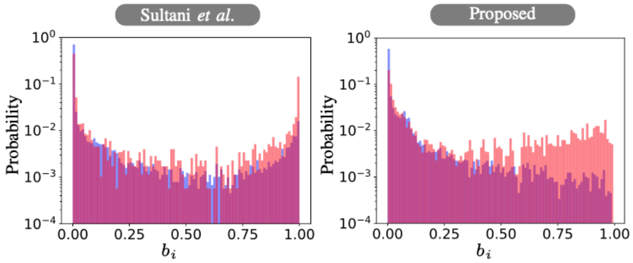

Figure 2: Comparison between the histograms describing the spread of $f(\mathcal{V}_u^i)$ scores in the unit interval, without/with the novel term. The left plot regards Sultani *et al.*'s loss, while the right plot provides the values for the loss function proposed in this paper. For visualization purposes, the vertical axis is plotted in log scale.

As stated above, the rationale of (4) is that less peaked distributions of scores will produce higher entropy values. Even if it is less frequent to obtain scores close to 0 and 1, the important point is to assure that only the most evidently *normal/abnormal* instances attain values close to the extremes, which is the key feature for selecting the instances that are iteratively added to the next generation of the learning sets, by means of the Bayesian framework.

3.2. Self-supervised Network

The SS model works at the segment level, and follows a classical classification paradigm. It receives C3D [15] feature vectors of 4,096 components describing video segments (16 frames/each) and predicts the *abnormality* of that segment, i.e., the likelihood of containing one abnormal event. This network has a 3-layer fully connected architecture, with 512 units in the first layer, 32 units in the second one and 1 unit in the output layer. We used ReLU [10] activation functions for the two first layers and a sigmoid activation in the output neuron. A binary cross-entropy loss drives the learning process of this model.

3.3. Bayesian Classifiers

In this work, the Bayesian classifiers enable to perceive - among the unsupervised data - the degree of belief for each classified instance. In a self-supervised fashion, the idea is that only the instances with the highest beliefs should be added to the next generation of the learning set, either for the WS and SS models. There are two Bayesian classifiers in our method: 1) the *weak* one (P_w), that receives the scores produces by the WS expert and filters the segments that should be used in the SS training; and 2) the *strong* counterpart, that receives the scores from the SS expert (at the segment level), and selects the videos that should be used in the next iteration of the WS model.

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

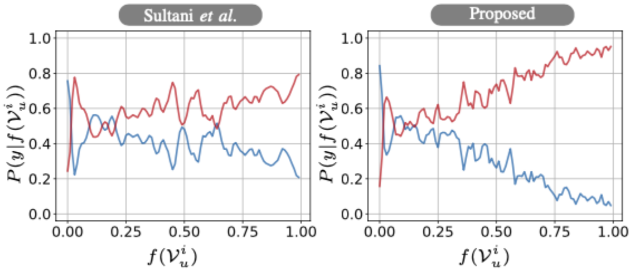

Figure 3: Comparison between the posteriors $P\big(y|f(\mathcal{V}^i)\big)$ obtained by the Bayesian classifier, when using the Sultani *et al.* formulation and ours.

where $y \in \{'a','n'\}$ represent the *abnormal/normal* classes. A Gaussian kernel density estimator was used to more accurately approximate the conditional densities $P\big(f(\mathcal{V}^i)|y\big)$, with Scott's rule [13] for bandwidth selection. Even though there is typically a significant imbalance between the number of *normal* and *abnormal* instances, we empirically adjusted the priors to $P('a') = P('n') = 0.5$.

Fig. 3 compares the posteriors per class, according to the Sultani *et al.*'s loss and ours, when the entropy term was also considered. The blue lines represent the *normal* events, and the red lines represents the *abnormality* scores. As it is notorious, the immediate effect of our proposal is to obtain a more evident separation between the degrees of belief for both classes near the extremes of the unit interval (particularly in the upper extreme).

Formally, the ith segment is selected for the next generation of the SS learning data according to the rule:

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

i.e., if the posterior for either the 'n'/'a' classes is higher than a threshold.

Regarding the *strong* Bayesian classifier, it should decide at the bag (video) level, receiving the video segments scores and classifying globally that video as *normal/abnormal*:

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

where $y \in \{'a','n'\}$ represent the *abnormal/normal* classes. Having initially considered the utilisation of simple fusion rules (such as max or $prod$) for obtaining the degree of belief per video, the poor levels of performance led us to the choice of a simple feed-forward classification module with four layers and $256 : 196 : 128 : 1$ architecture for estimating $P_s\big(y|f(\mathcal{B}^i)\big)$. This network was trained in a small part of our initial learning set (at iteration 0). Finally, and in way much similar to the weak Bayesian classifier, only the video instances that provide the extreme degrees of belief are selected to the next iteration of the learning set:

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

The τ_i values were adjusted empirically, according to the performance observed in a validation set.

4. Experiments and Discussion

4.1. Datasets, Baselines and Empirical Evaluation Protocol

Our experiments were conducted in two datasets: 1) concerned about the detection of *fights* in surveillance footage, we mined 1,000 videos (collected from *Youtube* and *LiveLeak*): 784 containing normal events, and the remaining 216 containing some *fight* segment. This dataset was manually annotated at the frame level; and 2) the UCF-Crime [14], composed of 1,900 surveillance videos, including *normal* scenes and 13 types of abnormalities, annotated at the video level (learning) and at the segment level (test). It is currently the largest freely available data source for abnormal events detection.

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

As baselines, five methods were considered to represent the state-of-the-art: Sultani *et al.* [14], Ravanbakhsh *et al.* [11], Chong and Tay [2], Hasan *et al.* [5] and Wang *et al.* [17], all described in section 2. Apart these, we also tested the performance attained by a quadratic kernel SVM working under the single-class paradigm, fed by the same C3D feature vectors as our WS and Sultani *et al.* models. The area under the curve (AUC), and the equal error rate (EER) and used as performance measures. Also, considering that abnormal events are rare, we also use the TPR@FPR=0.001 to report the typical performance values at very low error rates.

To guarantee a fair evaluation between our method and the state-of-the-art, considering that one premise of our work is that a set of unsupervised data is available, we disregarded the ground-truth labels of 50% of the learning sets,

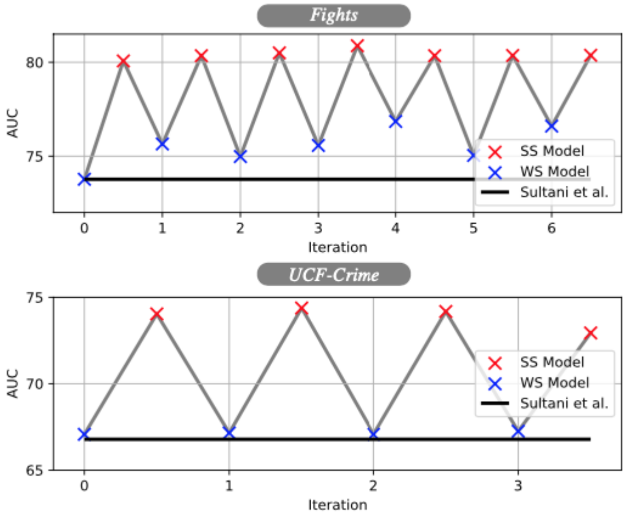

Figure 4: Improvements in validation performance of our frameworks with respect to the iteration, observed for the *Fights* and *UCF-Crime* sets. The red points denote the SS model and the blue points denotes the WS model. As baseline, the continuous black lines denote Sultani *et al.*'s performance.

and considered those subsets as *unlabelled*. Accordingly, these instances were also removed from the learning sets of the other methods. In practice, all methods used rigorously the same amount of annotated labels for earning purposes.

4.2. Results and Discussion

4.2.1 Individual Methods

The first observation was the poor performance that single-class techniques attain in this type of problem, in accordance to the observation drew by Sultani *et al.* In our view, this might be due to the extreme heterogeneity of the expected inputs, with these frameworks eventually requiring additional amounts of learning data to become minimally effective.

Regarding our proposal, Fig. 4 provides the evolution in performance the WS and SS models, according to the number of iterations. For both the *Fights* and *UCF-Crime* datasets, the SS attains its optimal point at the end of the third iteration, while the WS model obtained its maximum AUC value at the forth iteration, by a clear margin for the *Fights* set, and only by a residual difference for the *UCF-Crime*. Regarding the results in the latter set, the substantially harder task of finding simultaneously consistent patterns among 13 different types of abnormalities justifies the performance of the WS expert, which got results practically equal to the Sultani *et al.* method in all iterations.

Fig. 5 compares the ROC curves of all methods, in the *Fights* (left plot) and *UCF-Crime* (right plot) sets. Considering that normal events are rare, the zoomed-in plots in each plot provide the performance values for very low FAR values (TPR@FAR ¡ 0.01). Overall, our method attained clearly better results than any competitor, with exception of a small region in the performance space of *UCF-Crime* set, where Ravanbakhsh *et al.* outperformed any other (even here, our solution became the best one for FAR ¿ 0.006). Interestingly, the improvements in performance were more evident in the *UCF-Crime* set than in *Fights*, even considering that our method was designed having in mind the specific detection of fights. Sultani *et al.* was the runner up approach in most cases, which accords the previously reported results. However, the gap observed between the results reported here and in [14] enables to conclude about an extreme sensitivity of this technique to the amounts of data used for learning purpose. All the other methods were considered to completely fail in the detection of abnormal actions, at least with these amounts of learning data, and for heterogeneous environments as the considered here. In our view, the application of single-class learning techniques would be appropriate exclusively when providing learning data that features the exact same lighting conditions and camera pose as the test set.

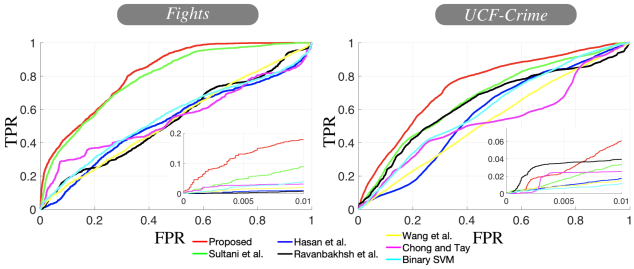

Figure 5: Comparison between the ROC curves obtained by the seven methods considered, for the *Fights* (left plot) and *UCF-Crime* (right plot) datasets. The inner plots provide the performance levels at low false acceptance rates.

4.2.2 Late Fusion Ensemble

The linear correlation between the responses produced by all the methods is shown in the left plot of Fig. 6. A 8 × 8 matrix is provided, describing the Person correlation values between ours WS/SS models, Sultani *et al.* (S), Hasan *et al.* (H), Ravanbakhsh *et al.* (R), Wang *et al.* (W), Chong and Tay (C) and SVM classification (V). The upper diagonal provides the values for *UCF-Crime*, while the *Fights* values are in the lower diagonal of the matrix. The levels of correlation in both sets were roughly the same, even though slightly higher values were observed for the *Fights* set. Also, the SVM classifier was - overall - the method that

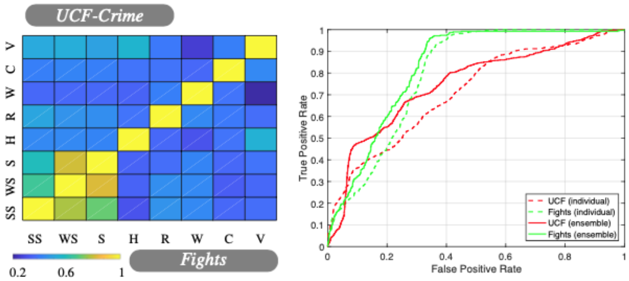

Figure 6: At left: linear correlation (Pearson coefficient) between the responses given by the methods analyzed. At right: improvements in performance for the *Fights* and *UCF-Crime* sets when using a Random Tree ensemble that fuses at the score level the output of the WS/SS/[14] experts (continuous lines) with respect to the best individual expert (dashed lines).

got the lowest average levels of correlation with respect to the remaining techniques.

Such relatively low levels of linear correlation motivated us to exploit the improvements in performance due to a late fusion scheme (Random Forest [1]). We created fifty bootstrap samples, each composed of 90% of the learning instances (drew with repetition) and created a decision tree for each sample. Each tree receives the scores provided by the three best performing algorithms (our SS+WS models plus [14]). Then, in inference time, the final score is given by the mean prediction among all the decision trees. The results are provided at the right plot in Fig. 6 and turn evident the improvement in effectiveness with respect to the best individual expert, attaining AUC values of about 0.819 (*Fights*) and 0.769 (*UCF-Crime*). The overall performance values are sumarized in Table 1 for all the methods tested individually and also for the ensemble.

Finally, as an ablation experiment, Fig. 7 enables to perceive the important role of the novel term added to the Sultani *et al.* [14]'s loss. Both plots compare the results observed when λ_3=0, i.e., when the entropy term is not used to span the scores in the unit interval (corresponding to Sultani *et al.*'s original formulation, in dashed lines). The continuous lines give the results when $\lambda_3 \in [0, \frac{\log(\text{ep}+1)}{30}]$ (ep: epoch). It is obvious that the entropy term improved the overall performance, both in the *Fights* and *UCF-Crime* sets. However, we note that improvements were particularly evident for moderate FAR values ([0.1, 0.6] interval), which is not the most useful region for the deployment of this kind of systems. We are currently working on the development of alternative solutions to extend the improvements in performance also for very low FAR values.

Method	**AUC**	**EER**	**TPR@FPR=0.001**
Fights			
Hasan *et al.* [5]	0.528 ± 0.003	0.466 ± 0.002	0.001 ± 9e^{-4}
Ravanbakhsh *et al.* [11]	0.533 ± 0.003	0.484 ± 0.002	0.002 ± 0.001
Wang *et al.* [17]	0.540 ± 0.002	0.475 ± 0.002	0.004± 0.001
Chong and Tay [2]	0.541± 0.003	0.480 ± 0.001	0.010 ± 0.003
Binary SVM Classifier	0.556 ± 0.003	0.443 ± 0.003	0.004 ± 0.002
Sultani *et al.* [14]	0.787 ± 0.002	0.294 ± 0.002	0.014 ± 0.002
Proposed Method	**0.819** ± 0.001	**0.284** ± 0.001	**0.046** ± 0.007
Fusion (SS+WS+[14])	**0.846** ± 0.001	**0.216** ± 0.001	**0.049** ± 0.001
UCF-Crime			
Hasan *et al.* [5]	0.573 ± 0.001	0.424 ± 0.001	0.003 ± 0.001
Ravanbakhsh *et al.* [11]	0.642 ± 0.001	0.376 ± 0.001	**0.008** ± 0.003
Wang *et al.* [17]	0.539 ± 0.001	0.472 ± 0.001	0.002 ± 0.001
Chong and Tay [2]	0.532 ± 0.002	0.484 ± 0.001	0.001 ± 7e^{-4}
Binary SVM Classifier	0.604 ± 0.001	0.441 ± 0.001	0.002 ± 9e^{-4}
Sultani *et al.* [14]	0.668 ± 0.001	0.375 ± 0.001	0.001 ± 8e^{-4}
Proposed Method	**0.744** ± 0.001	**0.305** ± 0.001	0.002 ± 0.001
Fusion (SS+WS+[14])	**0.769** ± 9e^{-4}	**0.266** ± 9e^{-4}	0.002 ± 7e^{-4}

Table 1: Performance summary of the proposed method with respect to the state-of-the-art. The final rows in each dataset provide the performance of an ensemble Random Forest that fuses at the score level the responses of our SS/WS models to [14].

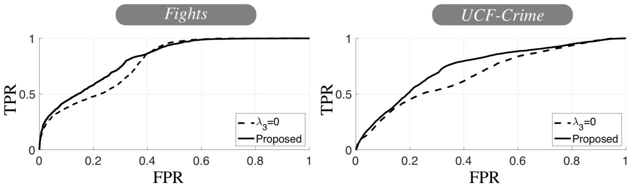

Figure 7: Ablation studies: results when λ_3=0, i.e., when the entropy term is not used to span the scores in the unit interval (dashed lines). The continuous lines correspond to the results when λ_3 is defined as given in Sec. 4.1.

5. Conclusions and Further Work

In this paper we described a self-supervised learning scheme composed of two experts (based in the weakly and self-supervised paradigms) feeding data to each other. Under this paradigm, we use two Bayesian frameworks to filter the input instances of the weakly and self-supervised experts, and iteratively augment the available learning data. Most importantly, we used a late fusion ensemble strategy (Random Forest of 50 decision trees) that fuses at the score level the outputs of the three best performing models tested. This ensemble outperforms the state-of-the-art in both datasets, decreasing the equal error rates from 15%

(*UCF-Crime*) to 24% (*Fights*). As an additional contribution, we announced the free availability of the *Fights* dataset, fully annotated at the frame level. The source code of the method and the datasets are freely available, and can be used to support further developments in the problems of human activity analysis/abnormal events detection.

Acknowledgements

This work is funded by the FEDER, under the PT2020 program and project POCI-01-0247-FEDER-033395. Also, this work is funded by FCT/MCTES through national funds and co-funded by EU funds under the project UIDB/EEA/50008/2020.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[2] Y. S. Chong and Y. H. Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*, pages 189–196. Springer, 2017.

[3] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[7] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[9] M. Moya and D. Hush. Network constraints and multi- objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996.

[10] V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines vinod nai. In *Proceedings of the IEEE international conference on machine learning*, pages 807–814, 2010.

[11] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017.

[12] P. Roy and G.-A. Bilodeau. Adversarially learned abnormal trajectory classifier. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 65–72. IEEE, 2019.

[13] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[14] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[16] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[17] T. Wang, M. Qiao, Z. Lin, C. Li, H. Snoussi, Z. Liu, and C. Choi. Generative neural networks for anomaly detection in crowded scenes. *IEEE Transactions on Information Forensics and Security*, 14(5):1390–1399, 2018.