Explainable Anomaly Detection in Videos Based on the Description of Atomic Actions

Andrei de Souza Inácio, Raphael Marinho Teixeira, Heitor Silvério Lopes
Bioinformatics and Computational Intelligence Laboratory – LABIC/CPGEI
Federal University of Technology Parana (UTFPR)
Emails: {andrei.inacioo, raphinhamt}@gmail.com
hslopes@utfpr.edu.br

***Abstract*—Anomaly detection in surveillance videos is an exhaustive and tedious task to be performed manually by humans. Many methods have been proposed to detect anomalous events by learning normal patterns and differentiate them from abnormal ones. However, these methods often suffer from false alarms, as human behaviors and environments can change over time. In addition, these methods fail to discriminate the types of anomalies that can occur, especially in anomalies performed by humans. This study presents an approach to detect anomalous events based on atomic action descriptions. It combines a tracking people method with atomic action detection and recognition network to understand video events and generate atomic descriptions. Besides detecting the anomalies, the proposed approach can also describe the anomalous action with human attributes in natural language. Anomalies are detected based on the generated descriptions of the scene. Experimental results show the effectiveness of our approach, presenting an average F1-Score of 87%.**

***Index Terms*—Anomaly detection; Deep Learning; Atomic video description**

I. Introduction

Automatically detecting events in videos is essential for video surveillance, which can be used to monitor and pre-alarm abnormal human behaviors or events in public and private spaces. Such an automatic system is also crucial in the security area to reduce the human effort required to monitor videos manually 24 hours a day [1], facilitate retrieval events of interest, and provide better efficiency in terms of person identification and recognition, help to solve crimes, and, in some cases, prevent them.

A challenging task in video surveillance is detecting anomalous events such as illegal activities or anomalous behaviors. Anomalous events in surveillance videos are generally defined as irregular or unexpected events that deviate from the normal ones [2]. The definition of an anomaly is strongly dependent on human-defined semantics and on the context where it happens. For example, the event "skating" inside an airport lounge can be considered an anomalous event while it is normal on outdoor squares or streets.

Existing methods for anomaly detection, in special those based on deep learning, are normally opaque. Usually, they provide only a binary decision regarding the presence or not of an anomaly.

However, a step beyond would be to explain why a given event is an anomaly. This is specially important in safety-critical applications, such as the surveillance of people in restricted areas. Such approach goes towards the Explainable AI (XAI) paradigm [3].

Atomic actions are simple activities or atomic body movements that can be described with few words such as walking, drinking, or holding an object and have the potential to become building blocks for more complex actions or activities. Thus, such information can be used to detect and explain activities performed by humans.

In this study, we describe an anomaly detection system for video streams based on human-comprehensible language. It consists of a hybrid system that combines the detection of atomic actions of multiple persons using the AIA network [4] with Yolo-v4 [5], a network specialized in extracting soft biometrics attributes, and a template-based generation method for the atomic actions description. A case-study is presented for video segments with normal and abnormal events, taken in airport lounges, where security is always an important issue.

The main contribution of this study is to demonstrate the feasibility of the proposed method for the anomaly detection task. Furthermore, such a method can be easily generalized to detect other anomalous events in different contexts.

This paper is structured as follows. Section II presents some related works in the field of anomaly detection. Section III describes the proposed method. Section IV shows the experimental results and discussion. Section V presents the conclusion and future works.

II. Related Work

Previous approaches for the anomaly detection task have used hand-crafted features such as motion and appearance [6]. In the last years, Deep Learning (DL) techniques have achieved promising results in anomaly detection by learning better features with superior discriminatory power for video and images representation [7], [8]. DL techniques have also been applied to solve different tasks in computer vision, including action recognition [9] person re-identification [10], age and gender recognition [11], and clothing segmentation [12], [13].

In Anomaly Detection task, traditional DL approaches usually employ a semi-supervised or unsupervised method by training a model to learn normal events and then detect anomalous events considered different from the normal ones [14].

In [15] is proposed an approach that uses Convolutional Autoencoder (CAE) to learn normal behaviours and, then, uses the trained model for anomaly detection in an one-class classification problem. [16] proposed an unsupervised deep learning approach known as ISTL for anomaly detection and localization for real-time video surveillance. It is based on autoencoder model combined with Convolutional Neural Network(CNN) and Convolutional LSTM (ConvLSTM) layers to automatically learn spatio and temporal features. In [17] is proposed a method to model the normal patterns of human movements in surveillance video for anomaly detection using dynamic skeleton features.

Unlike semi-supervised or unsupervised methods, and motivated by the insufficient discriminatory ability of the existing approaches and the complicated context and diverse different behaviors of people in videos, [18] proposed a method for detecting and locating anomalies in a supervised way. It is composed of two main modules: a human detection module, based on the Yolo network, to detect people from videos, and an anomaly detection module, which performs a binary classification and an action recognition task.

Since the events and the environment captured by surveillance cameras can change drastically over time, these approaches tend to produce high false alarm rates [19]. Moreover, although the good performance achieved by DL approaches for anomaly detection and recognition, they mainly focus on the detection accuracy aspect and fail to explain the detected anomalies [20]. Explaining abnormal events is essential for helping human observers quickly judge if they are false alarms or not [2]. Using an anomaly detection system that can explain abnormal events detected, unimportant events can be identified quickly, and human inspection can be avoided. Figure 1 shows an example of the ideal output generated by an anomaly detection and description system.

III. METHODS

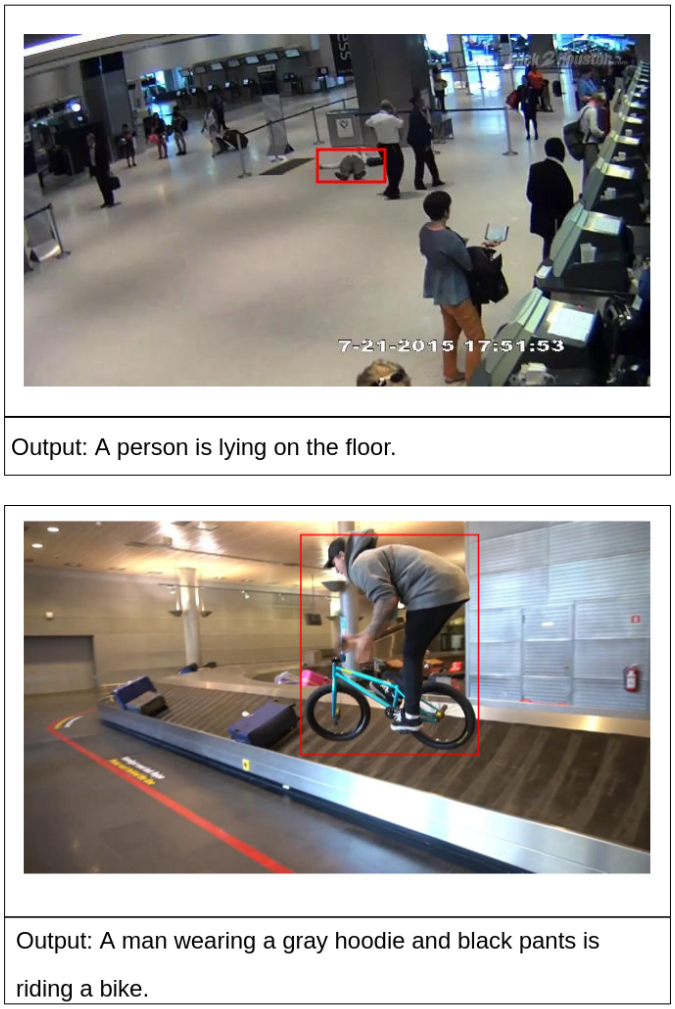
In this study, the proposed method, datailed in the following Sections, consists of four main components: Atomic Action Detection, Soft Biometrics Extraction, Atomic Action Description, and Anomaly Detection Procedure. Figure 2 shows the functional flowchart of the proposed method.

A. Atomic Action Detection

This module aims to detect and recognize the atomic actions of multiple people in videos. Given an input video, four frames per second were extracted, then the Deep SORT algorithm [21] (an improvement of the SORT algorithm [22]), integrated with Yolo-V3 [23] was used to track and detect people.

Then, the SlowFast Network [24] based on Resnet-101 [25], pre-trained on Kinetics-700 dataset [26], was used to extract features for each detected person.

Finally, we use the Asynchronous Interaction Aggregation (AIA) model [27], pre-trained on AVA dataset [28], to recognize the actions performed by each person. It has achieved the state of the art performance on action recognition recently.



Output: A person is lying on the floor.

Output: A man wearing a gray hoodie and black pants is riding a bike.

Figure 1. Examples of anomaly detection and its corresponding description.

The output of this module consists of a list of detected people with spatial annotation (bounding boxes), the actions detected for each person with the temporal annotation (beginning and end of the action), and a confidence score of each detected action.

B. Soft Biometrics Extraction

Soft biometrics traits, such as gender, age, clothes, and behaviors, are human characteristics that can help to distinguish and describe people.

We choose to extract gender information from facial attributes in this study, but other soft biometrics traits could be easily included, such as age and clothing information.

First, the YOLOV2-tiny, pre-trained on Face Detection Data Set and Benchmark (FDDB) [29], was used to detect the face from a person in a given bounding box. We use the pre-trained weights available at the YoloKerasFaceDetection project[1].

Then, a cropped face is used to detect the gender information. A neural network based on EfficientNet Network [30]

[1] https://github.com/abars/YoloKerasFaceDetection

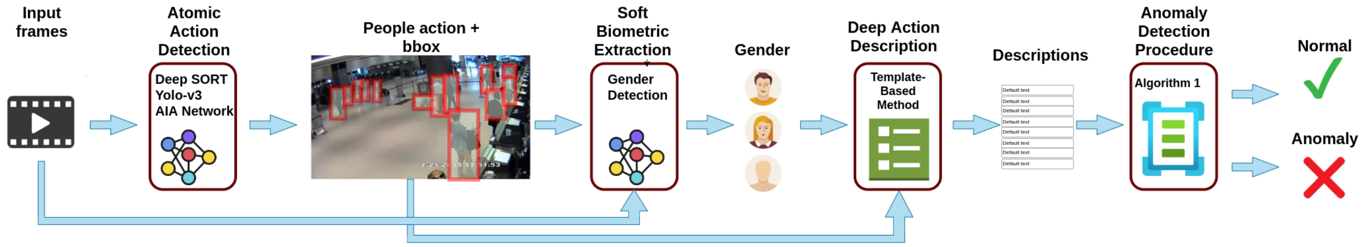


Figure 2. Overview of the proposed approach.

trained on the IMDB-Wiki dataset [31] was used for the gender classification task.

C. Atomic Action Description

The atomic action description consists of a method to generate a short sentence in natural language describing details of the person and the atomic action detected in a given video.

As some actions have human-object interaction, the Yolo-V3 network was also used to detect objects that overlap with the subject bounding box and may be essential to generate the description.

In our experiments, a template-based method, inspired by [32], was used to generate anomaly descriptions of surveillance videos in natural language. Based on the matched template, the natural language description of the detected anomalous event is generated. For example, a common anomaly description is represented by the template: "A {person | man | woman} is {riding | driving} a { object_name}". In our experiments, we have defined ten different templates to describe the perceived atomic actions.

D. Anomaly Detection Procedure

This module works at a semantic level, identifying whether anomalous events are happening in videos based on the description of the atomic actions provided by the previous module. Our experiments considered the following actions as possible anomalies: drive, dance, swim, kick something, play a musical instrument, shoot, fight or hit somebody, run or jog, jump or leap, and ride (car, bike, skateboard). Note that some of these actions are context-dependent and eventually need additional information to define as an anomaly (for example, in an airport, a person running late for the flight is a normal event, but a person running in a forbidden place can be an anomalous action). Collective actions are not addressed in this study, but they can be treated by the proposed method, by considering simultaneous action descriptions in the same time. Algorithm 1 presents the pseudocode for the proposed anomaly detection module.

Algorithm 1: Anomaly Detection Algorithm

```
Input: list of atomic action descriptions $l_{ad}$
Output: list of detected anomalous action descriptions $l_d$
Data: anomalous object-action mapping $l$.
foreach description in $l_{ad}$ do
    foreach action in $l$ do
        if description contains action then
            if action requires object description then
                if description contains object then
                    append(description) to the returning array $l_d$
            else
                append(description) to the returning array $l_d$
return $l_d$
```

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents experimental results performed to assess the feasibility of the proposed method in detecting anomalous events based on atomic actions descriptions.

The reported experiments were performed on a workstation with Intel Core-i7 8700 processor, 32GBytes RAM, and a Nvidia Titan-Xp GPU. The Tensorflow 2.2.0 and Keras 2.3.1 library were used to train and test the models described in Section III-A.

For the case study conducted in this work, a new dataset was created with airport surveillance videos extracted from the Youtube website.

Results were evaluated quantitatively, calculated from the F1-Score and accuracy metric, and qualitatively by visual inspection.

A. Dataset

The majority of most used datasets for the anomaly detection task, such as UCSD [33], and Avenue [34], focus on actors performing a simulated anomaly event, including walking in the wrong direction, loitering, or throwing objects, against an uncluttered or unaltered background. Also, some of them only provide frames in grayscale or have low resolution, which are insufficient to detect and describe important information for surveillance, such as the age, gender, or clothing color of a person. Therefore, although they are used for the anomaly detection task, they are not suitable for describing the abnormal events in surveillance videos in detail.

Thus, we create a new dataset for airport surveillance by collecting videos from the Youtube website, including normal and anomalous events in videos recorded by security systems and ordinary people using smartphones. The dataset has a total duration of 33.5 minutes. This dataset, named UTFPR-AASD [2] (Airport Anomaly Surveillance Dataset), includes 70 videos, 45 of which contain anomalies. Besides the “normal” class, five different anomaly classes were considered in these videos, including:

1) Suspicious actions: running or jumping;
2) Unusual actions: kicking or dancing;
3) Violent actions: fighting or shooting;
4) Fainting;
5) Driving and riding.

Each video was manually annotated with the type of anomaly and temporal information (beginning and ending of the anomaly). Although not useful for our study, temporal information has been included and will be considered in future works related to the anomaly localization task.

Figure 3 shows some examples of abnormal events included in the dataset. We can note that the dataset has anomalous actions involving objects (Figure 3A), human-objects interactions (Figure 3B) and 3C), and isolated human actions (Figure 3D). The videos were recorded in real-world situations with different resolutions and illumination conditions, making the problem even more challenging.

Notice that the anomalous events flagged in these videos are context-dependent. In our case, they are considered anomalies within an airport lounge, but could, eventually, be considered normal in other scenarios.

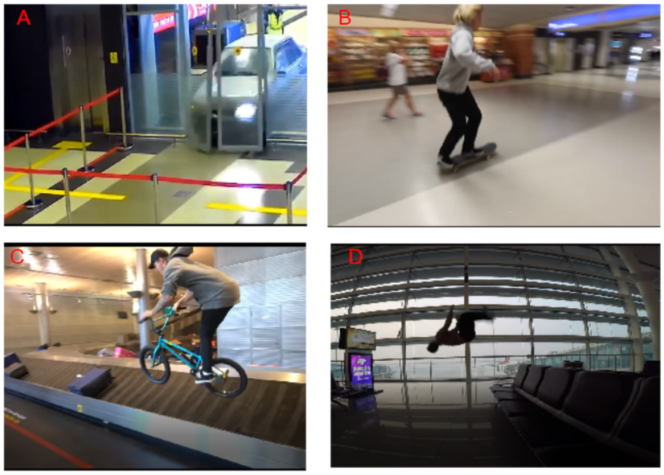


Figure 3. Examples of anomalies in the dataset. A) a person is driving a car. B) A man is skateboarding. C) A man is riding a bike. D) A person is jumping and vaulting over the seats.

B. Quantitative results

Table I shows the performance of the proposed method. The average F1-score achieved was 87%, indicating that anomalies based on atomic actions can be satisfactorily detected by this method.

The confusion matrix, presented in Figure 4, shows the correctly detected and wrongly detected anomalies. It was observed that the proposed method could detected 84% of the anomalies presented in the dataset.

Table I
Anomaly Detection Performance.

	Precision	Recall	F1-Score
Normal	0.77	0.92	0.84
Anomaly	0.95	0.84	0.89
Average	0.86	0.88	0.87

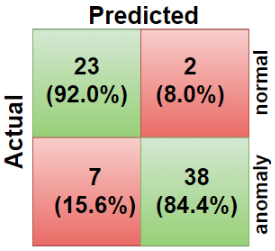


Figure 4. Confusion Matrix.

A possible advantage of the proposed method, compared to traditional anomaly detection methods, is that it goes beyond detecting, since it can describe anomalous events in human-comprehensible language. Also, when new normal situations are detected, the method can be easily adapted to avoid false alarms (false positives).

Table II presents the average accuracy considering the normal class and all anomaly classes (see Section IV-A). Notice that the proposed method works well with an unbalanced dataset and can detect anomalous events satisfactorily.

Table II
Abnormal event detection accuracy by class.

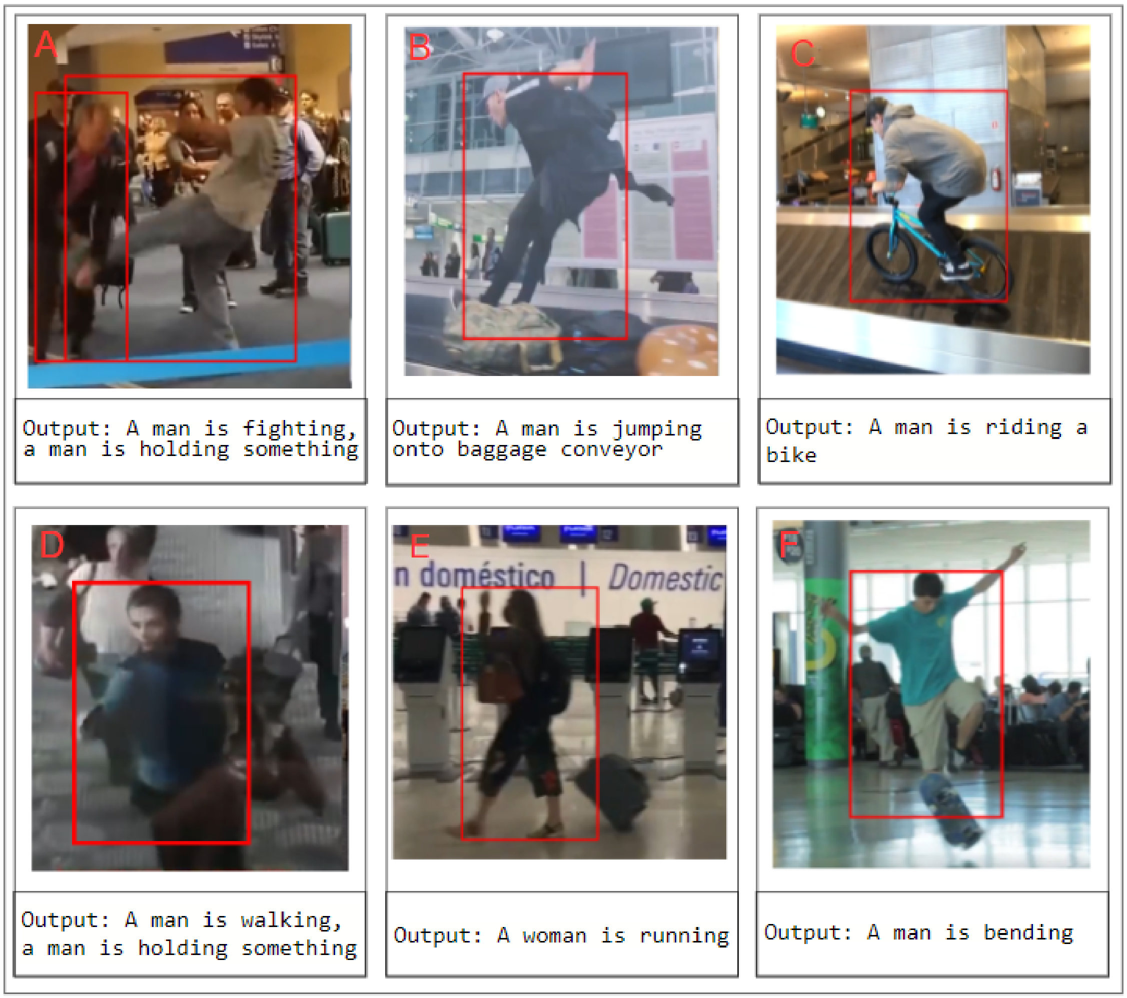
Anomaly	# Videos	Detected	Accuracy
Normal	25	23	92,00%
Suspicious action	10	10	100,00%
Unusual actions	12	10	83,33%
Ride or Drive	13	11	84,61%
Fainting	2	1	50,00%
Violent actions	8	6	75,00%
Total	70	59	87,71%

C. Qualitative results

Figure 5 shows qualitative results of some events detected as anomalies and also false positive results, as follows:

- Figure 5A, 5B, 5C: show some samples of anomalies with the atomic action description provided by the proposed method. Notice that it is able to detect different types of anomalous events, including action with human-object interaction.
- Figure 5D: presents a false negative example of a man shooting with a fire gun. The method could not detect the action and neither the weapon presented in the frames

[2] Dataset available on http://labic.utfpr.edu.br/datasets.html



```
Output: A man is fighting,
a man is holding something
```

```
Output: A man is jumping
onto baggage conveyor
```

```
Output: A man is riding a
bike
```

```
Output: A man is walking,
a man is holding something
```

```
Output: A woman is running
```

```
Output: A man is bending
```

Figure 5. Example of anomaly detection and description provided by the proposed method. A) B) and C) are True Positive examples; D) and F) are a False Negative examples; and E) is a False Positive example.

due to the low resolution of video frames and illumination variation. This highlights the difficulty of detecting anomalies in surveillance videos under these conditions. Also, some preprocessing steps in the videos could be provided to better capture these situations.

- Figure 5E: presents a false alarm (false positive), by wrongly indicating the action running. It was observed that the atomic action detection process detects only in some frames the action running wrongly with low confidence. It indicates that the proposed method could be improved to discard random some actions detected at random in the video. Actually, running is an ambiguous action that can be considered normal in some situations (for example, a person can be running late for the flight time).
- Figure 5F: presents a false negative instance. The proposed method detects that a man is fighting, which was right in some frames, but it could not detect the skateboarding action, indicating that the action detection network still can be improved, since it may fail in some situations.

V. CONCLUSION

Anomaly detection in surveillance videos is a non-trivial problem since anomalous events are highly dependent on human concepts that rarely occur compared to normal activities. Consequently, most strategies to automatically detect anomalies focus on learning normal events and detecting deviations from the normality. Due to the fact that both, human behaviors and environments may change over time, the traditional anomaly detection systems may generate high false alarm rates.

This paper have presented a method to detect anomalous events in surveillance videos based on atomic action descriptions. It combines the use of an atomic action detection and recognition network with an action description approach to generate a sequence of sentences for a video. Then anomalous events are detected based on these generated descriptions. Furthermore, the proposed method is a step towards a more explainable artificial intelligence, since it provides a kind of explanation of what are the anomalies detected.

Based on the results presented here, the proposed method was able to satisfactorily detect anomalies in surveillance videos for specific contexts, based on atomic actions descriptions. The proposed system can be easily adapted to different contexts by modifying the rules and the atomic actions in the anomaly detection module.

This is an ongoing work, and some improvements in the proposed system will be done in the near future. Among the issues to be addressed are: (1) lack of anomaly detection when the actions take place far from the point where the scene was filmed, and when there are occlusions of the actors involved in the scene; and (2) difficulty in identifying collective actions into atomic actions, when several people are the actors of the scene. Also, future works focus, also, on enriching the description, by detecting more human details in the descriptions, such as age/gender and type/color of clothes.

Acknowledgment

A. S. Inácio thanks UNIEDU/FUMDES - PÓS GRADUAÇÃO for the scholarship and Federal Institute of Santa Catarina for the support; H.S. Lopes thanks CNPq for the research grants no. 311785/2019-0, and Fundação Araucária for the research grant PRONEX 042/2018. Special thanks to NVIDIA Corp. for the donation of the Titan-Xp GPU boards used in this study.

References

[1] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 79, pp. 1–17, 2021.

[2] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proc. of IEEE International Conference on Computer Vision*, 2017, pp. 3619–3627.

[3] E. H. F. Hussain, R. Hussain, "Explainable Artificial Intelligence (XAI): An Engineering Perspective," *arXiv preprint*, vol. 2101.03613v1, pp. 1–17, 2021.

[4] J. Tang, J. Xia, X. Mu, B. Pang, and C. Lu, "Asynchronous interaction aggregation for action detection," in *Proc. of European Conference on Computer Vision (ECCV)*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer, 2020, pp. 71–87.

[5] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint*, vol. arXiv:2004.10934, 2020.

[6] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," in *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Press, 2018.

[7] R. de Paula Monteiro and C. J. A. Bastos Filho, "Feature extraction using convolutional neural networks for anomaly detection," in *Anais do 14^{o} Congresso Brasileiro de Inteligência Computacional*. Curitiba, PR: ABRICOM, 2019, pp. 1–8.

[8] M. Ribeiro, M. Romero, A. Lazzaretti, and H. S. Lopes, "Learning spatio-temporal features for detecting anomalies in videos using convolutional autoencoder," in *Anais do 14^{o} Congresso Brasileiro de Inteligência Computacional*. Curitiba, PR: ABRICOM, 2019, pp. 1–8.

[9] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "MoViNets: Mobile video networks for efficient video recognition," *arXiv preprint*, vol. arXiv:2103.11511, 2021.

[10] M. Ye and P. C. Yuen, "Purifynet: A robust person re-identification model with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2655–2666, 2020.

[11] C.-Y. Hsu, L.-E. Lin, and C. H. Lin, "Age and gender recognition with random occluded data augmentation on facial images," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 631–11 653, 2021.

[12] A. S. Inácio and H. S. Lopes, "Epynet: Efficient pyramidal network for clothing segmentation," *IEEE Access*, vol. 8, pp. 187 882–187 892, 2020.

[13] A. de Souza Inácio, A. Brilhador, and H. S. Lopes, "Semantic segmentation of clothes in the context of soft biometrics using deep learning methods," in *Anais do 14^{o} Congresso Brasileiro de Inteligência Computacional*. Curitiba, PR: ABRICOM, 2019, pp. 1–7.

[14] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: a survey," *arXiv preprint*, vol. arXiv:1901.03407, 2019.

[15] M. Ribeiro, A. E. Lazzaretti, and H. S. Lopes, "A study of deep convolutional auto-encoders for anomaly detection in videos," *Pattern Recognition Letters*, vol. 105, pp. 13–22, 2018.

[16] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393–402, 2020.

[17] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Press, 2019.

[18] M. Gong, H. Zeng, Y. Xie, H. Li, and Z. Tang, "Local distinguishability aggrandizing network for human anomaly detection," *Neural Networks*, vol. 122, pp. 364–373, 2020.

[19] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Press, 2018, pp. 6479–6488.

[20] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 2021.

[21] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. of IEEE International Conference on Image Processing (ICIP)*. IEEE Press, 2017, pp. 3645–3649.

[22] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. of IEEE International Conference on Image Processing (ICIP)*. IEEE Press, 2016, pp. 3464–3468.

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint*, vol. arXiv:1804.02767, 2018.

[24] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. of IEEE/CVF International Conference on Computer Vision*. IEEE Press, 2019, pp. 6202–6211.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Press, 2016, pp. 770–778.

[26] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint*, vol. arXiv:1907.06987, 2019.

[27] J. Tang, J. Xia, X. Mu, B. Pang, and C. Lu, "Asynchronous interaction aggregation for action detection," in *Proc. of European Conference on Computer Vision*, 2020, pp. 71–87.

[28] C. Gu, C. Sun, D. A. Ross, C. Vondrick *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[29] V. Jain and E. Learned-Miller, "FDDB: a benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[30] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. of International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[31] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 144–157, 2018.

[32] S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Natural language description of surveillance events," in *Information Technology and Applied Mathematics*, P. Chandra, D. Gin, E. Li, S. Kar, and D. K. Jana, Eds. Singapore: Springer, 2019, pp. 141–151.

[33] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Press, 2010, pp. 1975–1981.

[34] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. of IEEE International Conference on Computer Vision*. IEEE Press, 2013, pp. 2720–2727.