Review of Physical Aggression Detection Techniques in Video Using Explainable Artificial Intelligence

Pablo Negre[1(✉)], Ricardo S. Alonso[2,3], Javier Prieto[1], Angélica González Arrieta[1], and Juan M. Corchado[1]

[1] Grupo de Investigación BISITE, Universidad de Salamanca, 37006 Salamanca, Spain
{pablo.negre,javierp,angelica,corchado}@usal.es, corchado@air-institute.com

[2] AIR Insitute, Paseo de Belen 9A, 47011 Valladolid, Spain

[3] UNIR (International University of La Rioja), Av. de la Paz, 137, Logrono, Spain
ricardoserafin.alonso@unir.net

Abstract. Physical aggressions are a problem that affects our society throughout the world. This study is a review of the work done so far on the physical detection of aggressions in video using XAI techniques (XAI). In the Related Work section, other reviews have been presented that compile physical aggression detection, XAI techniques, and the combination of both: physical aggression detection using XAI techniques. The state of the art section, compiles papers on video assault detection, an explanation of XAI and the different ways of categorising and grouping existing techniques, and a review of papers dealing with video assault detection using XAI. The search methodology used consisted of using Google Scholar, reviewing articles with a publication date of no more than two years, with high quality and with a theme as close as possible to the subject matter of the article. As far as it has been possible to find, no works have been found on the detection physical aggressions or actions using XAI techniques; however, some works have been found in the context of object detection using XAI techniques. This opens the possibility of creating a research line on the detection of physical aggressions in video using XAI.

Keywords: Physical aggression detection · Explainable artificial intelligence · Computer vision

1 Introduction

Physical assaults are a serious and widespread problem in society, affecting individuals all over the world. Real-time detection of these situations can be a key factor in protecting and saving the lives of vulnerable people, thus ensuring the physical safety of the population. Supporting this assertion, according to a World

P. Novais et al. (Eds.): ISAmI 2023, LNNS 770, pp. 53–62, 2023.
https://doi.org/10.1007/978-3-031-43461-7_6

Health Organisation report, it is estimated that one in three women worldwide have been subjected to either physical and/or sexual intimate partner violence or non-partner sexual violence in their lifetime [23].

The use of security cameras has largely been used to provide evidence of a crime that can be used later to detect the culprits or serve as evidence for those affected in the eyes of insurance companies, police or judges. While it is a matter of debate, because of the balance between citizen security and individual rights, China is a country that has opted for drastic surveillance through the massive installation of security cameras; which aim, among other things, to detect and prevent crime [12]. The development of technologies in sectors such as electronics, big data, artificial intelligence (AI), IoT [19], etc., has made it possible to obtain and process a huge amount of data from the environment around us. This rise of big data and AI technologies [9] allows for real-time video analysis. This has opened up the possibility of developing security camera analysis for real-time crime detection through the use of computer vision algorithms. The development of optimal detection algorithms would allow for automatic real-time detection, which would make it possible to take immediate action such as calling the authorities [24]. All this added to the growing need for the generated algorithms to provide explainable results [8], so that the prediction is reliable, makes this a research topic of great interest.
The working methodology consisted of using Google Scholar, a well-known meta-search engine for articles. Articles with a publication date of no more than two years and whose subject matter was as similar as possible to the detection of video aggression using explainable artificial intelligence (XAI) were selected.

This work is a review of the state of art in the detection of physical aggressions in video using AI algorithms and from the approach of XAI. Section 2 and 3 present, respectively, related work and state of the art on the detection of video aggression using XAI techniques. Finally, Sect. 4 presents the conclusions as well as future lines of work.

2 Existing Reviews on the Detection of Video Aggression Using XAI Techniques

Section 2 provides an overview of related work on the detection of aggressions using computer vision techniques. It includes a presentation of previous research on aggression detection and an overview of reviews focusing on XAI. However, there is a lack of reviews discussing the detection of video aggression using XAI techniques.

There are multiple articles on the review of aggression detection techniques. A large part of them divide the detection process into three main parts: the detection method, the feature extraction and object detection, and the properties of the dataset and video used for model training [25]. Within the detection methods, methods based on surface features have to be designed manually and have higher demands on the video quality. On the other hand, deep learning based methods obtain from training with the set of videos to obtain the best

detection method [30]. Regarding the dataset used as training videos, it has been studied how the same detection method behaves with different databases of aggression videos. This is important since the methods used may vary in their detection quality depending on the type of video in question [18]. Other reviews studies have also compiled studies on aggression detection in crowded areas, where ensuring security becomes more complex [28].

There are numerous reviews that discuss explainable and trustworth AI. They are mostly based on the great development of AI in today's society and how its use has made it a concern to be able to understand its results (explainable) and to be able to trust them (trustworth) [10]. The review article [20], exposes the challenges that still exist in the field of XAI. Among these challenges are the lack of universal definitions, standards and measures for the explainability of AI systems, the lack of balance between explainability and performance, and the difficulty of making deep learning models explainable.

After a thorough search of the literature, no reviews have been found that jointly address video-based physical assault detection together with XAI techniques. Therefore, there is an interesting research opportunity to explore this area further, as both video-based assault detection and XAI are currently highly relevant topics.

3 State of the Art on the Detection of Video Aggression Using XAI Techniques

Section 3 is a review of the state of the art in aggression detection using XAI techniques. Section 3.1 collects the latest studies related to the detection of aggressions in video by using computer vision techniques. Section 3.2 outlines the need for XAI, different approaches and classifications of XAI methods. Section 3.3 consists of a compilation of recent papers focusing on object detection using XAI techniques, due to the lack of work on both aggression and action detection in general.

The methodology used consisted of the selection of quality articles, with a publication date of less than two years and with the closest proximity to the subject of this work, the detection of video aggression using XAI techniques. Google Scholar was used to select the articles.

3.1 Detection of Aggressions Using Computer Vision Techniques

The detection of aggression has been a growing research topic given its broad applications in public and industrial environments [27]. However, manual detection of violent behaviors is very labor-intensive and has a high personnel and time cost, which can result in loss of lives [16]. On the other hand, aggression detection is complex due to its ambiguity, as it is difficult to differentiate aggression from other types of actions [7], as well as factors such as changing lighting, complex background, and low resolution, among others [27]. All of this makes it difficult to describe aggression through simple rules or classifiers using trivial algorithms [7].

In [2], a traditional method has been used for the detection of aggression, which are primarily based on the analysis of visual and motion features present in the video motion features present in the video. Feature extraction is performed using the STIP and MoSIFT descriptors. A visual word vocabulary is created using the k-means algorithm, and visual word histograms are created using the feature descriptors and the created vocabulary. Finally, using an SVM classifier, events are classified as aggressive or non-aggressive. Both the use of STIP and MoSIFT had an accuracy close to 90%. The dataset used contained videos of hockey games, whose content is usually more homogeneous.

In addition to the use of traditional methods, a large number of works perform video aggression detection with methods based on deep neural networks (DNN). The multimodal function has been used for aggression detection with a focus on four multimodal fusion methods using audio, video, and text modalities, obtaining accuracies of 86%. The proposed architecture is the use of DNN (deep neural networks), which are effective not only in prediction tasks but also in feature extraction and dimensionality reduction [7]. A Multi-Scale Spatio-Temporal Network (MSTN) has also been used [31]. This consists of a spatiotemporal module that extracts features from the original video. Then, double pooling layers are used, which apply filters to the extracted features to create different scales of temporal information. Then, two modules STB (focused on object location and shape) and LTB (focused on object velocity and motion) are used to extract violence-specific features. Finally, the Trans module merges the results obtained from STS and LTB, obtaining a violence/non-violence result. Videos of actual assaults have been used, as well as videos of field hockey games and action scenes from movies. Results with an accuracy of 90.25% are obtained using the RWF-2000 dataset [3] (containing real assaults), obtaining a higher accuracy than with the use of algorithms used in other works. In [27], a convolutional neural network (CNN) model is used for pre-processing, searching for important objects such as humans and vehicles. Then it proceeds to the feature extraction phase to gather information and characteristics of the sequence. After that, features are extracted through another convolutional neural network (CNN) that is concatenated with the high-level features of the Darknet-19 (CNN) model, forming the final feature map. In the third step, an LSTM network is used that obtains sequence features and learns them for violence detection. Results with high accuracy are obtained for the three datasets used: 98.21%, 98%, and 74% on the Violent Flow, Hockey Fight, and Surveillance Fight datasets, respectively. Another study that uses convolutional neural networks (CNN) is [6]. CNNs have been compared with the models: MobileNet model (96.6%), AlexNet (88.99%), VGG (96.49%), GoogleNet (94.99%). The proposed MobileNet model has shown outstanding performance in the perspective of accuracy, loss, and computation time on the hockey fight dataset. In article [16] CNNs have also been used. The VGG-16 model (of CNN type) is used for feature extraction, the results of which feed a ConvLSTM for final classification. This architecture has obtained high accuracies in six different dataframes with values above 98% in five out of six of them.

All in all, there are several positive points in the works mentioned above. The use of feature extraction in the detection process allows relevant information to be captured. The creation of public databases containing examples of physical assaults has facilitated the training and evaluation of models. Also, the improvement in accuracy with the use of CNN has led to a clear improvement in accuracy compared to other types of algorithms. On the negative side, the wide variety of scenarios, lighting, recording quality, as well as the use of videos of non-real assaults (films, hockey games...) complicates the detection of real assaults.

3.2 Explainable Artificial Intelligence Techniques

XAI is growing rapidly due to the increasing need for connection between machines and human behaviour [21], developing interpretable algorithms that provide understandable explanations to humans of decisions made by the model [8]. Model explainability is not only targeted at AI experts but also at model end users and students [26]. In [1] and arguing against XAI, it is cited how some healthcare personnel (as end users of predictive model output) preferred factual information from trustworthy sources, rather than a complete understanding of how the information was generated as well as greater difficulty in detecting erroneous predictions. However, in [14] it is exposed how trustworthy AI requires XAI explainability, which is but a requirement to reach a high state of trustworthiness. It was found in [26] that while different explanatory models can give interpretable motivations, among them there are key differences in the results obtained by each. It was also noted how the chosen method of explainability influences the significance of the features more than the underlying model and the data. XAI methods can be classified according to several [5] factors: intrinsic or post hoc (depending on whether the model is interpretable for simplicity or not), model-specific or model-agnostic (depending on whether the XAI model is independent of the model used or not) and local or global (whether one seeks to understand the general behavior of the model or of a particular prediction). On the other hand, there are three approaches to obtain an explanatory AI using a deep neural network. The first is by visualizing or analyzing the internal behavior of DNNs (without modifying the existing structure). The second is to add an explanatory module to the DNN to analyze the input and generate an explanation of why a certain prediction has been produced. The third consists of creating a new interpretable model (from scratch) to make decisions through the explainable [22] representations.

In [17] an explainable-by-design model is proposed for object detection based on a combination of semantic and visual properties, so it would fall under the classification of intrinsic explainable model, since its results are understandable per se. In this case the XAI approach cannot be selected since the underlying model is not a DNN. Saliency maps are algorithms used to provide visual explanations for neural networks based on computer vision. These algorithms show the regions relevant to the decision that the AI model made (so they correspond to the first of the explanatory AI approaches). In [8] the Grad-CAM method and

its variants are used to produce high quality saliency maps (obtaining the best results Grad-CAM and XGrad-CAM).

In conclusion, the growing need for algorithms whose results are explainable has made XAI a growing field of study. The use of visual tools such as saliency maps to highlight important areas of an image makes it easier to understand the results. A better understanding of the outcome of the algorithms without the need for advanced technical knowledge is also an advance. On the other hand, the performance and computational complexity of XAI processes can still be an issue. While the clustering of XAI algorithms into different groups is an advance in their variety and development, there is a lack of metrics to assess the effectiveness of the applied method.

3.3 Explainable Artificial Intelligence in Aggresion Detection

Section 3.3 exposes existing articles that deal with explainable and trustworthy AI in the detection of aggressions. As will be shown below, as far as we know, there are no papers dealing with this topic, so we will present articles on explainable and trustworthy AI in the detection of actions (as a field of study closer to the detection of aggressions).

To the best of our knowledge, there is no paper addressing the detection of physical aggression by AI techniques using video cameras, from the point of view of XAI. Since physical aggression detection using video cameras is a growing research topic, as discussed throughout this paper, a whole avenue of research remains open to address XAI techniques that ensure that the detection of aggression by the model is understandable. This would allow trustworthy detection that avoids both confusion with other contact actions (e.g., hugging) and detection by people with different physical and clothing characteristics. Since no papers were found on XAI in aggression detection, we searched for papers on XAI in action detection, which is a broader set in which aggression detection is found. However, no papers could be found on XAI applied to action detection either. All in all, it was decided to look for papers on XAI in object detection, as this is a much more studied field within computer vision. In [15] the detection of autonomous vehicles has been studied using the "KITTI road dataset", using as detection algorithms: ResNet-18, ResNet-50 and SegNet; and as XAI methods Grad-CAM, Saliency maps and Intermediate DL model layer analysis. While the use of various XAI techniques provides greater clarity on the 'black box' operation of the model, the explainability of the model is not fully resolved; it is proposed to integrate NLP into the XAI system to provide explanations through the use of text. In [13], 10 different object classes of the autonomous driving environment are detected using the BDD100K dataset, the YOLOv4 model was used as the detection algorithm and RISE (which generates saliency maps) was used as the XAI tool; it is argued that the distance between vehicles and other objects should be taken into account in future work for correct detection. In [29], the MSCOCO and Flickr30K datasets (containing images from multiple different categories) are used and a CNN is used for object detection together with an RNN that associates the detected object with a word describing it,

making the model intrinsically explainable. In [11] multiple datasets have been used: COCO, KITTI, BDD, and OpenImages; YOLOv3 has been used as object detection algorithm and Grad-CAM has been integrated into the model. In [4] L-CRP, an XAI algorithm based on heat maps for the most used state-of-the-art object detection algorithms (UNet, DeepLab & Yolo), has been created using multiple public datasets such as: CityScapes, Pascal VOC 2012 and MS COCO 2017; demonstrating the fidelity of L-CRP.

In summary, as far as is known, there is no work on the detection of aggressions or actions using XAI algorithms. As far as object detection is concerned, the works carried out share the datasets used and widely used XAI algorithms, obtaining good results and facilitating comparison between them (Table 1).

Table 1. Summary table of the state of the art of object detection using XAI techniques.

Cite	Detection category	Dataset	Detection algorithms	XAI algorithms
[15]	Vehicle detection	KITTI road	ResNet-18, ResNet-50 and SegNet	Grad-CAM, Saliency maps and DL model middle layer analysis
[13]	Vehicle detection	BDD100K	YOLOv4	RISE (saliency maps)
[29]	Multiple categories	MSCOCO and Flickr30K	CNN + RNN	Intrinsically explainable
[11]	Multiple categories	COCO, KITTI, BDD and OpenImages	YOLOv3	Grad-CAM integrado en el modelo
[4]	Multiple categories	CityScapes, Pascal VOC 2012 and MS COCO 2017	UNet, DeepLab and Yolo	L-CRP

4 Conclusions and Future Work

Physical assaults are a serious and widespread reality in our society around the world. Real-time detection can be a key factor in ensuring the physical safety of the population in dangerous situations.

With the development of technology, it is possible to process information on a large scale, and for this to be done in real time. This opens up the possibility that physical aggressions can be detected in real time [24] with an algorithm that can send an alarm to the competent authorities. In addition to the great potential this offers, there is also a growing need for AI algorithms to be explainable for developers and users [8], making the combination of real-time physical assault detection and the use of AI a research topic of great interest.

This work is a review of the state of art in the detection of physical aggressions in video using AI algorithms and from the approach of XAI. Section 2 presents

related work on video assault detection using XAI, highlighting the importance of the chosen dataset on the accuracy of the algorithm, the complexity of detection in crowded areas and the need for greater understanding of the algorithm results.

Section 3 first discusses on Sect. 3.1 the state of the art on the detection of physical aggression in video using AI. Some positive points mentioned about the studied works are: the better results with the use of feature extraction, the creation of public databases tagged with aggression/non-aggression videos and the improvement in accuracy that the use of CNN has brought about. As negative points, the variety in video features and the use of non-real assault videos (movies, hockey games...), which make it difficult to detect real assaults. Section 3.2 presents work on XAI techniques, which is a growing field of study. Positive points to highlight in the works studied are the use of visual tools for the interpretation of the result, as well as the non-necessity of being a technical person to understand the result. On the downside, performance and computational complexity remain issues. In Sect. 3.3, to the best of our knowledge there are no works that bring together these fields of research, where the physical detection of aggressions through video by means of XAI is dealt with. While there are a number of papers that address XAI in object detection [11,15], where positive results have been presented with widely used XAI algorithms.

As future lines of research and as a continuation of this first review of the state of the art on the detection of physical aggressions in video by using XAI techniques, a systematic study will be carried out that will be relevant to structure the wide variety of works in which different algorithms and databases have been used, both in the detection of physical aggressions by using video and XAI techniques. Subsequently, it is proposed that, based on the information gathered in this article and this systematic study, algorithms for the detection of physical aggression using video XAI techniques, whose accuracy is at least as high as that of current algorithms, should be developed.

Acknowledgements. This work was supported by the project "XAI - Sistemas Inteligentes Auto Explicativos creados con Módulos de Mezcla de Expertos", ID SA082P20: financed by Junta Castilla y León, Consejería de Educación, and FEDER fund.

References

1. Amann, J., et al.: To explain or not to explain?-artificial intelligence explainability in clinical decision support systems. PLOS Digit. Health **1**(2), e0000016 (2022)
2. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011. LNCS, vol. 6855, pp. 332–339. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23678-5_39
3. Cheng, M., Cai, K., Li, M.: RWF-2000: an open large scale video database for violence detection. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4183–4190. IEEE (2021)

4. Dreyer, M., Achtibat, R., Wiegand, T., Samek, W., Lapuschkin, S.: Revealing hidden context bias in segmentation and object detection through concept-specific explanations. arXiv preprint arXiv:2211.11426 (2022)
5. Hall, S.W., Sakzad, A., Choo, K.K.R.: Explainable artificial intelligence for digital forensics. Wiley Interdisc. Rev. Forensic Sci. **4**(2), e1434 (2022)
6. Hussain, T., Iqbal, A., Yang, B., Hussain, A.: Real time violence detection in surveillance videos using convolutional neural networks. Multimedia Tools Appl. **81**(26), 38151–38173 (2022)
7. Jaafar, N., Lachiri, Z.: Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. Expert Syst. Appl. **211**, 118523 (2023). https://doi.org/10.1016/j.eswa.2022.118523. https://www.sciencedirect.com/science/article/pii/S0957417422016013
8. Karim, M.M., Li, Y., Qin, R.: Toward explainable artificial intelligence for early anticipation of traffic accidents. Transp. Res. Rec. **2676**(6), 743–755 (2022)
9. Katta, P., Kandasamy, K., Raj, R.S.P., Subramanian, R., Perumal, C.: Regression based performance analysis and fault detection in induction motors by using deep learning technique. ADCAIJ: Adv. Distrib. Comput. Artif. Intell. J. **11**(3), 349–365 (2023). https://doi.org/10.14201/adcaij.28435. https://revistas.usal.es/cinco/index.php/2255-2863/article/view/28435
10. Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A.: Trustworthy artificial intelligence: a review. ACM Comput. Surv. (CSUR) **55**(2), 1–38 (2022)
11. Kirchknopf, A., Slijepcevic, D., Wunderlich, I., Breiter, M., Traxler, J., Zeppelzauer, M.: Explaining yolo: leveraging grad-cam to explain object detections. arXiv preprint arXiv:2211.12108 (2022)
12. Kostka, G., Steinacker, L., Meckel, M.: Between security and convenience: facial recognition technology in the eyes of citizens in china, Germany, the United Kingdom, and the United States. Public Underst. Sci. **30**(6), 671–690 (2021)
13. Li, Y., et al.: A deep learning-based hybrid framework for object detection and recognition in autonomous driving. IEEE Access **8**, 194228–194239 (2020)
14. Ma, J., et al.: Towards trustworthy AI in dentistry. J. Dent. Res. **101**(11), 1263–1268 (2022)
15. Mankodiya, H., Jadav, D., Gupta, R., Tanwar, S., Hong, W.C., Sharma, R.: OD-XAI: explainable AI-based semantic object detection for autonomous vehicles. Appl. Sci. **12**(11), 5310 (2022)
16. Mugunga, I., Dong, J., Rigall, E., Guo, S., Madessa, A.H., Nawaz, H.S.: A frame-based feature model for violence detection from surveillance cameras using convlstm network. In: 2021 6th International Conference on Image, Vision and Computing (ICIVC), pp. 55–60 (2021). https://doi.org/10.1109/ICIVC52351.2021.9526948
17. Olszewska, J.I.: Snakes in trees: an explainable artificial intelligence approach for automatic object detection and recognition. In: ICAART (3), pp. 996–1002 (2022)
18. Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M.: State-of-the-art violence detection techniques in video surveillance security systems: a systematic review. PeerJ Comput. Sci. **8**, e920 (2022)
19. Qader Kheder, M., Aree Ali, M.: Iot-based vision techniques in autonomous driving: a review. ADCAIJ: Adv. Distrib. Comput. Artif. Intell. J. **11**(3), 367–394 (2023). https://revistas.usal.es/cinco/index.php/2255-2863/article/view/28821
20. Rawal, A., McCoy, J., Rawat, D.B., Sadler, B.M., Amant, R.S.: Recent advances in trustworthy explainable artificial intelligence: status, challenges, and perspectives. IEEE Trans. Artif. Intell. **3**(6), 852–866 (2021)

21. Rodríguez Oconitrillo, L.R., Vargas, J.J., Camacho, A., Burgos, Á., Corchado, J.M.: RYEL: an experimental study in the behavioral response of judges using a novel technique for acquiring higher-order thinking based on explainable artificial intelligence and case-based reasoning. Electronics **10**(12), 1500 (2021)
22. Sakai, A., et al.: Medical professional enhancement using explainable artificial intelligence in fetal cardiac ultrasound screening. Biomedicines **10**(3), 551 (2022)
23. de la Salud, O.M.: Violence against women (2021). https://www.who.int/news-room/fact-sheets/detail/violence-against-women
24. Shah, N., Bhagat, N., Shah, M.: Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Vis. Comput. Ind. Biomed. Art **4**(1), 1–14 (2021). https://doi.org/10.1186/s42492-021-00075-z
25. Siddique, M., et al.: State-of-the-art violence detection techniques: a review. Asian J. Res. Comput. Sci. 29–42 (2022)
26. Swamy, V., Radmehr, B., Krco, N., Marras, M., Käser, T.: Evaluating the explainers: black-box explainable machine learning for student success prediction in MOOCs. arXiv preprint arXiv:2207.00551 (2022)
27. Ullah, F.U.M., et al.: An intelligent system for complex violence pattern analysis and detection. Int. J. Intell. Syst. **37**(12), 10400–10422 (2022). https://doi.org/10.1002/int.22537. https://onlinelibrary.wiley.com/doi/abs/10.1002/int.22537
28. Ullah, F.U.M., Obaidat, M.S., Ullah, A., Muhammad, K., Hijji, M., Baik, S.W.: A comprehensive review on vision-based violence detection in surveillance videos. ACM Comput. Surv. **55**(10), 1–44 (2023)
29. Wu, T., Song, X.: Towards interpretable object detection by unfolding latent structures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6033–6043 (2019)
30. Yao, H., Hu, X.: A survey of video violence detection. Cyber-Phys. Syst. **9**(1), 1–24 (2023)
31. Zhou, W., Min, X., Zhao, Y., Pang, Y., Yi, J.: A multi-scale spatio-temporal network for violence behavior detection. IEEE Tran. Biometrics Behav. Identity Sci. **5**, 266–276 (2023)