





Evaluating Methods for Violence Classification and Firearm Detection in Indoor CCTV Environment

Arnaldo V. Barros da Silva   [Universidade Federal do Agreste de Pernambuco | arnaldovitorbarros@gmail.com]

Luis F. Alves Pereira  [Universidade Federal do Agreste de Pernambuco|luis-filipe.pereira@ufape.edu.br]

 Universidade Federal do Agreste de Pernambuco, Avenida Bom Pastor, Garanhuns, 55292-270, Pernambuco, Brazil

Received: 26 March 2023 • **Accepted:** 29 July 2024 • **Published:** 05 October 2024

Abstract The adoption of security systems based on computer vision for violence detection has the potential to significantly improve safety in various public and private properties. However, developing these systems can be extremely challenging. We can choose to use classification models to identify violence in images or also use object detection models to identify firearms, which may indicate robbery. Additionally, when developing such systems focused on private environments, we encounter specific challenges, such as obtaining appropriate datasets to train the algorithms. Many publicly available datasets for violence detection consist of outdoor images, with elements such as streets and cars, which do not adequately reflect the nuances and unique characteristics of private properties. In this work, we evaluate both learned and handcrafted features to classify videos as 'violence' or 'non-violence' across a variety of datasets, including a new dataset composed exclusively of closed-circuit television (CCTV) images. Additionally, we propose a new dataset for firearm detection in CCTV images and conduct some experiments using YoloV8. In this way, we hope to provide a clearer insight into the possible decisions when developing a security system for indoor environments.

Keywords: Violence Detection, Firearm Detection, Image Features, Deep Learning

1 Introduction

As the number of surveillance cameras installed at indoor environments worldwide increases, the urgency for real-time video-content analysis also grows. Once supervising multiple monitors during long hours is an unsuited task for human agents, many computer vision algorithms have been proposed during the last decade for detecting abnormal and potentially dangerous situations, such as: (i) people disobedience to virtual fence [Delgado *et al.*, 2014; Chen *et al.*, 2012]; (ii) loitering [Coşar *et al.*, 2016; Arroyo *et al.*, 2015]; (iii) crowd panic [Krausz and Bauckhage, 2012; Zhang *et al.*, 2019]; (iv) seniors falling [Lu *et al.*, 2018; Rougier *et al.*, 2011]; and others.

In the security sector, systems for detecting assaults in real-time can allow the development of automatic alerts informing potential risk situations directly to security agents. For example, detection of firearms [Grega *et al.*, 2016; Olmos *et al.*, 2018] can suggest an assault in progress. However, there are many more characteristics to consider, such as people in a panic, running, surrendering, fighting, or dragging furniture. Unfortunately, according to our current knowledge, there are not many available datasets that only include indoor footage with violence that possess the mentioned characteristics above.

Previous works on violence detection were built upon video clips of fights extracted from: (i) action movies [Nievas *et al.*, 2011]; (ii) matches of the National Hockey League (NHL)[Nievas *et al.*, 2011], (iii) real-world crowds [Hassner *et al.*, 2012; Setti *et al.*, 2017]; (iv) fighting sports

[Soomro *et al.*, 2012]; (v) real-world pedestrians [Blunsden and Fisher, 2010]; or (vi) real-world criminals [Sultani *et al.*, 2018; Cheng *et al.*, 2020]. None of the (i)-(v) datasets are suited for training methods to detect indoor violence. Furthermore, current datasets based on real-world criminal actions (vi) also contains outdoor scenes involving cars, streets, and other elements that introduces unnecessary features to the design of indoor violence detection systems. Thus, we propose a new benchmark for detection of real-world Assaults at Indoor Environments: the AIE dataset¹. Our dataset is composed of 700 video clips of surveillance cameras installed at bank correspondents, convenience stores, retail stores, and others. There are video clips with only regular customers, and other ones with attacks where criminals use guns occluded from the footage, melee weapons, handguns, rifles, machine guns, or no guns at all. In addition, as mentioned earlier, detecting firearms can be crucial to identifying ongoing violent events. With that in mind, we have created the Firearm Detection at Indoor Environments (FDIE) dataset², consisting of 2213 selected images annotated with bounding boxes highlighting different types of firearms.

We not only proposed these two new datasets but also conducted experiments with them. To understand the challenges in training a Deep Learning (DL) model on the detection data, we turned our focus to one of the most popular detection frameworks, YOLO (You Only Look Once) [Redmon *et al.*, 2016]. Developed by Joseph Redmon and others, YOLO was first introduced at CVPR 2016. This framework revolution-

¹Download at: <https://www.kaggle.com/arnaldovitor/aie-dataset>

²Download at: <https://www.kaggle.com/arnaldovitor/fdie-dataset>

ized object detection by introducing a real-time, end-to-end approach capable of performing detection tasks with a single pass through the network. Unlike previous methods that used sliding windows or two-stage approaches (where the first stage detected regions with possible objects and the second performed classification in those regions), YOLO simplified the output by using only regression to predict detection results.

As for the classification data, we compared classical computer vision architectures with new DL models. In classical computer vision, solutions for detecting violence in images often utilized handcrafted (HC) features based on optical flow [Gao *et al.*, 2016], appearance [Chen and Hauptmann, 2009], or acceleration [Deniz *et al.*, 2014]. These features were then encoded into numerical vectors, with each vector being assigned a label of '*violence*' or '*non-violence*'. Subsequently, machine learning techniques such as Support Vector Machine (SVM) [Hearst *et al.*, 1998] and Random Forest [Breiman, 2001] were employed to classify new images.

One of the most popular classical methods for video classification is MoSIFT [Chen and Hauptmann, 2009], which is based on the extraction of keypoints using the Scale-Invariant Feature Transform (SIFT) [Lindeberg, 2012]. These keypoints represent pixels in an image that are more distinctive than others. Among the extracted keypoints, those with optical flow greater than a certain threshold are selected. These selected keypoints are then associated with a descriptor composed of the concatenation of SIFT descriptors and Optical Flow Histogram (HoF) [Van Gool, 2008] descriptors. These descriptors are then used together with Bag of Visual Words (BoVW) [Csurka *et al.*, 2004] to create a representation of the images through histograms indicating the occurrences of certain characteristics. Using MoSIFT, accuracy rates were reported higher than other more modern classical methods [Zhou *et al.*, 2017] and even surpassing methods based on DL [Traoré and Akhloufi, 2020], reporting accuracy rates close to 94% in detecting violence in scenes of hockey matches and 89% in fights in crowds.

Although HC features have shown promising initial results, they have gradually been surpassed by DL based algorithms. The utilization of learned (LN) features has stood out in various areas such as image classification [Lu and Weng, 2007], object detection [Zou *et al.*, 2019], image segmentation [Garcia-Garcia *et al.*, 2017]. These features not only demonstrate high accuracy but also eliminate the complex task of designing specific feature extractors for images. This progress is enabled by the convolutional layers of Convolutional Neural Networks (CNNs), which extract relevant patterns through specialized sliding filters. In the realm of violent scene detection, the application of LN features has led to precision rates close to 100% in movies and hockey matches [Zhou *et al.*, 2017; Soliman *et al.*, 2019a; Keçeli and Kaya, 2017].

Despite its advantages, DL methods suffer from a lack of interpretability, exhibiting behavior similar to a black box [Lipton, 2018]. Interpretability is a crucial aspect for reliable intelligent systems, especially in high-risk scenarios such as violence detection. One way to shed light on the behavior of a DL model may include the Grad-CAM [Selvaraju *et al.*,

2017]. This method is model-agnostic, meaning it can be used across a wide range of CNNs without the need to modify their architectures. It operates by calculating the gradient of the last convolutional layer with respect to the CNN output, generating a heatmap that indicates the influence of each region in the generated prediction. Recent studies evaluating various eXplainable AI (XAI) techniques in diverse domains consistently found that Grad-CAM outperformed other methods such as LIME and SHAP, providing more reliable and robust visual interpretations of CNN decisions [Cian *et al.*, 2020; Varam *et al.*, 2023; Wei *et al.*, 2022].

Although LN features have become more popular than HC features, several researchers have invested efforts to determine whether it is possible to entirely disregard the classical approach: *i)* In a study investigating the skin cancer detection problem, it was found that classical solutions sometimes outperformed DL-based methods [Saba, 2021]. *ii)* A similar comparison was conducted in the context of pedestrian gender classification [Antipov *et al.*, 2015]. Their study demonstrated that both approaches had similar performance for small and homogeneous datasets, while LN features were more accurate in more complex data scenarios. *iii)* An extensive comparison between various classical and Deep Learning-based methods across a broad domain of images, ranging from butterfly species classification to cancer detection, revealed instances where HC features were more suitable [Nanni *et al.*, 2017].

In this study, we examine various aspects encountered in the development of violence detection systems. We focus specifically on the challenges associated with firearm detection and analyze several video datasets used for violence classification, each with unique characteristics. Instead of merely considering accuracy rates, we adopt more comprehensive approaches, such as visual explanation analysis. This approach aims to provide valuable insights that can contribute to enhancing violence detection methods.

The article is structured as follows: Section 2 introduces the AIE and FDIE datasets; Section 3 describes the experiments and parameters used for video classification; Section 4 details the experiments and parameters employed for firearm detection; Section 5 presents the obtained results and discussions; and finally, Section 6 concludes the paper.

2 Proposed datasets

2.1 AIE Dataset

The proposed dataset for detecting Assaults at Indoor Environments (AIE) comprises 700 clips of surveillance cameras published online, mostly at LiveLeak. Clips of 5 to 10 seconds limited to only ten clips per video were extracted from each footage to ensure data diversity. We also blurred faces from the images to ensure people's privacy. Finally, labels of '*violence*', '*non-violence*' and '*alert*' were assigned to each video clip. The '*alert*' label can be considered a subcategory of '*violence*' and was applied to 52 videos in the dataset where violence occurs in a way that is difficult to identify, even for a human observer, such as when a person has a occluded weapon.

Figure 1 shows the distribution of clips in the AIE dataset with respect to the time length (a), image resolution (b), environment (c), and weapon used (d). It is shown that most of the benchmark samples are between 9 and 10 seconds in length, and they have a resolution of 1280×720 . Furthermore, 70% of the videos in our benchmark depict environments of high occurrence of violent assaults: (i) Brazilian lottery houses - which are banking correspondents - and (ii) convenience stores worldwide. Besides that, in 26.75% of the violent clips, no fire gun was imaged.

Footage examples of the AIE benchmark are illustrated in Figure 2 that shows attacks using a handgun (a), and a rifle (b). Finally, the characteristics of the state-of-the-art benchmarks are compared with our AIE dataset in Table 1.

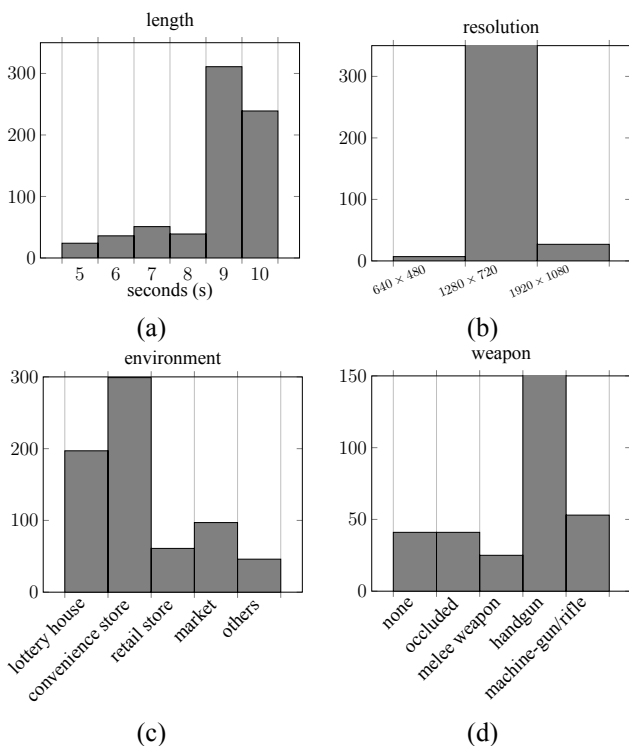


Figure 1. Distribution of clips in the AIE dataset with respect to the time length (a), image resolution (b), indoor environment (c), and the weapon used in the assaults (d).

2.2 FDIE Dataset

The proposed dataset for firearm detection, called the Firearm Detection at Indoor Environments (FDIE) Dataset, comprises 2213 frames extracted from the AIE dataset, with a total of 2312 annotated firearms in YOLO format. This dataset presents a significant challenge due to the statistics of the annotated bounding boxes, with an average of 5005 pixels, a minimum of 82 pixels, a maximum of 59673 pixels, and a standard deviation of 7387 pixels. These statistics indicate a wide variation in the size of the bounding boxes, which can impact the effectiveness of detection models and require robust approaches to handle the diversity of sizes and shapes of firearms present in the dataset. Examples of firearms that can be difficult to detect are illustrated in Figure 3 and finally, in Table 2 some of the most popular datasets are compared with the proposed one.



(a)



(b)

Figure 2. Examples of violent clips in the proposed dataset where criminals attacked with a handgun (a), and a rifle (b).

3 Video classification experiments

To maintain a fair comparison between HC and LN features we selected a representative method for each of them and evaluate their performances on different datasets. We also generated explanations of image classifications using Grad-CAM [Selvaraju et al., 2017] and BoVW-CAM [da Silva and Alves Pereira, 2022]. The BoVW-CAM is a method that provides visual explanations similar to Grad-CAM, but for HC features based on keypoints, as in the case of MoSIFT.

3.1 Selected datasets

To carry out the experiments, in addition to the proposed AIE dataset, another three sets were used: (i) the Hockey Fight dataset [Nievas et al., 2011], which contains 1000 videos extracted from National Hockey League (NHL) matches, manually labeled as 'fight' or 'non-fight'; (ii) the Violent Flows dataset [Hassner et al., 2012], which comprise violent and non-violent crowd scenes in real-world footage collected from YouTube and LiveLeak. The total of 246 videos include street protests, street fights, and stadium, among others; (iii) the RWF-2000 dataset [Cheng et al., 2020], which is composed of 2000 videos captured by surveillance cameras in real scenes and labeled as 'violence' and 'non-violence'. These datasets have been widely used in several recent works, such as Vijeikis et al. [2022] and Kang et al. [2021]

3.2 Handcrafted feature extractor

We used the MoSIFT [Chen and Hauptmann, 2009] technique to obtain a set of feature vectors for each input frame. This processing is described by the Algorithm 1. First, SIFT keypoints [Lindeberg, 2012] are computed for each frame input to find regions of interest. Then, a MoSIFT vector of size 256 is created by concatenating SIFT [Lindeberg, 2012] and HOF [Van Gool, 2008] descriptors for those regions of interest which have optical flow greater than a threshold ϵ .

The feature extractor is combined with the BoVW [Csurka et al., 2004] technique, which involves creating new video

Table 1. Characteristics of the most used datasets in the literature for automatic violence detection.

dataset	# violent	# non-violent	hours length	resource	release year
Behave Video [Blunsden and Fisher, 2010]	19	144	2.5	real-world outdoor	2010
Hollywood [Nievas et al., 2011]	100	100	0.08	movies	2011
Hockey [Nievas et al., 2011]	500	500	0.44	sports	2011
Violent Flows [Hassner et al., 2012]	123	123	0.8	real-world outdoor	2012
UCF101 [Soomro et al., 2012]	276	13044	27	sports	2012
UCF-Crime [Sultani et al., 2018]	950	950	127.8	real-world indoor and outdoor	2019
Real Life Violence Situations [Soliman et al., 2019b]	1000	1000	2.92	real-world indoor and outdoor	2019
Surveillance Camera Fight [Akti et al., 2019]	150	150	0.18	real-world indoor and outdoor	2019
CCTV-Fights [Perez et al., 2019]	1000	1000	17.68	real-world indoor and outdoor	2019
RWF-2000 [Cheng et al., 2020]	1000	1000	2.8	real-world indoor and outdoor	2020
AIE	400	300	1.83	real-world indoor	2023

representations through histograms. To generate these histograms, several steps are necessary. Firstly, a subset of the descriptors must be grouped using a clustering algorithm such as KMeans [Ahmed et al., 2020], which produces centroids known as visual words. Secondly, a dictionary of visual words is created using this set of visual words, and all descriptors extracted from the frames of a new video are associated with the visual word closest to them. Finally, the histogram that describes the video is generated by computing the frequency of occurrence for each visual word across its frames.

Algorithm 1: The handcrafted feature extractor

Input: $frame, next_frame, \epsilon$
Output: $hand_features$
 $hand_features \leftarrow \square$
 $keypoints \leftarrow SIFT(frame)$
for each $kp \in keypoints$ **do**
 if $opticalFlow(frame, next_frame, kp.position) > \epsilon$
 then
 $hof \leftarrow HOF(frame, next_frame, kp.position)$
 $mosift \leftarrow cat(hof, kp.descriptor)$
 $hand_features.add(mosift)$
 end
end

3.3 Learned feature extractor

We fine-tuned a pre-trained VGG-19 [Simonyan and Zisserman, 2014] model on the ImageNet [Deng et al., 2009]. However, unlike MoSIFT, which produces a single feature vector for each video, this architecture only classifies one image. To address this, we implemented a voting system where



Figure 3. Examples of firearms that are difficult to detect due to their dimensions.

a video is classified based on the most commonly assigned class among its frames.

3.4 Experimental parameters

We divided the dataset in two ways: to train the method based on LN features, 70% of the data were used. To train the HC

Table 2. Characteristics of datasets in the literature for firearm detection.

Dataset	# images	firearm type	resource	release year
Granada [Olmos <i>et al.</i> , 2018]	3000	handguns only	non-real world	2018
Monash Guns [Lim <i>et al.</i> , 2021]	3459	handguns only	partially real-world indoor and outdoor	2021
Shenzen [Qi <i>et al.</i> , 2021]	51889	various	partially real-world indoor and outdoor	2021
FDIE	2213	various	real-world indoor	2024

features method, the same previous training partition was divided into two of the same size, one to build the dictionary of visual words of the BoVW method and another to train the classifier. The other partitions in both approaches maintained the proportions, 10% for validation and 20% for testing. For the AIE dataset, undersampling was required to balance the data. We removed 52 examples from the 'alert' subclass and, randomly, 48 examples from the 'violence' class. This resulted in a total of 300 images for each of the 'violence' and 'non-violence' classes.

In relation to the hyperparameters of each method, the HC features-based method employed the K-Means clustering algorithm with 256 visual words to construct the dictionary. The threshold ϵ in Algorithm 1 was 0.5. The Fully Connected Network used as a classifier was trained for 200 epochs using the Adam optimization algorithm and the loss function was the Binary Cross-entropy. On the other hand, the employed in the LN feature-based method included 200 training epochs, the use of the SGD optimization algorithm, and the Categorical Cross-entropy loss function.

4 Object detection experiments

4.1 Model description

The selected object detection model for our experiments is YOLOv8 [Jocher *et al.*, 2023], which builds upon the architecture of YOLOv5 [Jocher, 2020]. YOLOv8 features an evolved backbone known as the C2f module, a modification of the CSPLayer from YOLOv5. This Cross-Stage Partial Bottleneck integrates high-level features with contextual information, significantly enhancing detection accuracy. Furthermore, state-of-the-art performance has Mean Average Precision (mAP) above 90% for firearm detection tasks [Khin and Htaik, 2024; Rastogi and Varshney, 2024].

We conducted our experiments by fine-tuning three variations of YOLOv8, all pre-trained on the COCO dataset [Lin *et al.*, 2014]. The primary difference between them lies in the number of trainable parameters: YOLOv8n with 3.2 million parameters, YOLOv8s with 11.2 million parameters, and YOLOv8m with 25.9 million parameters.

4.2 Training setup

The dataset was divided into training, validation, and test sets in proportions of 70%, 10%, and 20%, respectively. We ensured that frames from the same video were not included in different subsets, preventing any data leakage. The models were trained for 100 epochs following this strategy: during the first 10 epochs, we applied mosaic data augmentation [Hao and Zhili, 2020] to enhance the model's generalization

ability. In the subsequent epochs, we used more traditional augmentations such as horizontal and vertical flips, as well as changes in hue and saturation.

5 Results and Discussions

5.1 Classification accuracy

Table 3 shows the false-positive rates (FPR), false-negative rates (FNR), and accuracy rates (ACC) obtained in all video classification datasets evaluated in this work using their respective subsets of the test. It could be expected that LN features would always be better and this was not the case, HC features reported better accuracy rates in RWF-2000 and AIE, which are datasets with violence in the real world and therefore more challenging. AIE was the most challenging dataset for both approaches.

It is worth noting that other studies have achieved even higher accuracy rates, hitting 90% on the RWF-2000 dataset [Mohammadi and Nazerfard, 2023; Hachiuma *et al.*, 2023], surpassing our investigation's results. Nevertheless, this doesn't devalue our work. The differences in performance rates we've observed indicate that deep learning models might face challenges in grasping crucial characteristics, ones that classical methods can identify. This suggests a promising avenue for future research: combining these two distinct approaches.

Table 3. False-positive rate, false-negative rate, and accuracy rate obtained by the method based on HC and LN features in Hockey, Violent Flows, RWF-2000, and AIE sets.

Dataset	Feature	FPR	FNR	ACC
Hockey	Handcrafted	15.7%	10.8%	88.2%
Hockey	Learned	18.5%	4.3%	89.5%
Violent Flows	Handcrafted	14.8%	34.7%	76.9%
Violent Flows	Learned	14.6%	7.2%	88%
RWF-2000	Handcrafted	28.9%	30.1%	69.3%
RWF-2000	Learned	48.3%	30.1%	60.2%
AIE	Handcrafted	50.0%	33.3%	58.3%
AIE	Learned	61.6%	43.3%	47.5%

5.2 Venn Diagram

Accuracy alone cannot adequately determine the superiority of one method over another. To ensure a valid comparison, it is essential that all videos correctly classified by the method with lower accuracy be encompassed within the set of videos

correctly classified by the method with higher accuracy. An effective way to illustrate this issue is through a Venn Diagram comparing the predictions of the methods. Upon observing the Figure 4, it becomes apparent that in none of the datasets demonstrates complete overlap between the methods. Furthermore, the AIE dataset emerges as the most challenging. Even through the combination of predictions from both methods, a error rate of 22.50% persists.

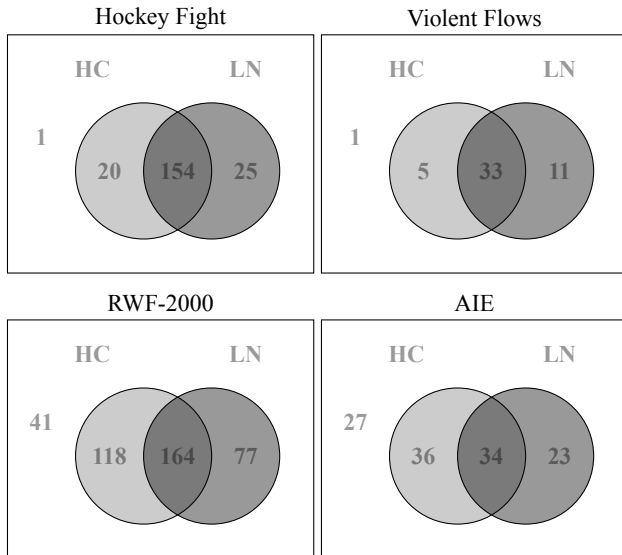


Figure 4. Number of videos classified correctly by handcrafted features (HC) and by learned features (LN).

5.3 Visualization

We generated visual explanations for the 'non-violent' (Figure 5) and 'violence' (Figure 6) class via Grad-CAM [Selvaraju et al., 2017] for the LN features based method and via BoVW-CAM [da Silva and Alves Pereira, 2022] the HC method. From these images, it is possible to see that both methods focus on different aspects to generate their classifications. The HC method evaluates a larger area of objects classified in the scene, in addition to that, it also focuses on objects that are in motion. The LN method focuses on image regions that are more difficult to be understood by human agents, but that still maximize accuracy.

5.4 Dice Score

To understand how distinct the focus regions of importance are between the HC and LN features, we transform the visual explanations of the Grad-CAM and BoVW-CAM into binary images for the entire test subset of each dataset to calculate the Dice Score (DS) between them [Dice, 1945]. The findings can be visualized in Table 4. It can be observed that the results, in general, were considerably low, suggesting a notable disparity between the characteristics each method focuses on. An interesting aspect is that the most challenging datasets (AIE and RWF-2000) showed a lower average DS when compared to the others.

Table 4. Average (Avg.) and standard deviation (Std.) obtained by calculating the Dice Score (DS) between the views generated by Grad-CAM and BoVW-CAM.

Dataset	Avg. of DS	Std. of DS
Hockey	0.411	0.194
Violent Flows	0.482	0.292
RWF-2000	0.361	0.248
AIE	0.162	0.200

5.5 Object detection results

Table 5 presents the precision (P), recall (R), and mAP obtained with the three models generated in this study on the FDIE dataset. The model that achieved the best results was YOLOv8s, demonstrating a precision exceeding 70%, which is promising considering a real-world environment. However, the results presented are still not satisfactory for high-risk scenarios. Our experiments suggest that to achieve robust performance with this data, it would be necessary to develop a specialized architecture due to the high variance in the size of the bounding boxes of the objects.

Table 5. Precision (P), recall (R), and mAP results for three YOLOv8 model variations on the FDIE dataset.

Model	P	R	mAP
YOLOv8n	0.699	0.318	0.383
YOLOv8s	0.730	0.395	0.455
YOLOv8m	0.690	0.320	0.378

6 Conclusion

In this work, we conducted experiments focused on firearm detection and video violence classification, and we also proposed two new datasets containing images of indoor environments for these tasks. In the classification domain, while the literature indicates that the most modern solutions are based on learned features, studies have suggested that handcrafted features capture different aspects of the image that can be advantageous for classification. We demonstrated that, even with generally higher accuracy rates, learned features cannot fully replace handcrafted features, as there are images that only the classical method classifies correctly. We also explored the divergence between the areas of interest of each method in the image domain using Grad-CAM and BoVW-CAM, and observed that the most relevant regions for each method are quite distinct.

For detection, we selected a modern object detection architecture that has shown promising results in firearm detection in the literature. However, we found that detecting firearms in CCTV images within our new dataset is significantly more challenging. Through these experiments, we aim to provide an initial pipeline and sufficient data to advance research in the development of intelligent systems for identifying ongoing assaults.

Grad-CAM

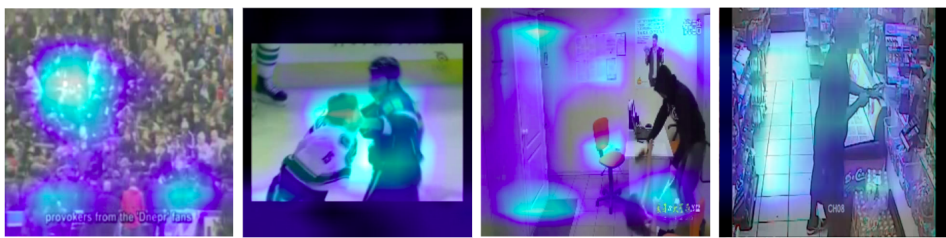


BoVW-CAM



Figure 5. Visualization for "non-violence" class with Grad-CAM and BoVW-CAM methods.

Grad-CAM



BoVW-CAM

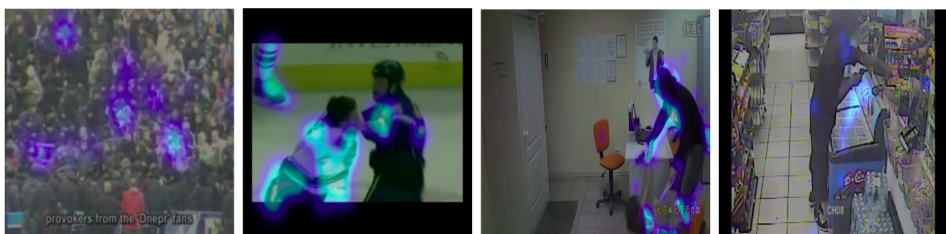


Figure 6. Visualization for "violence" class with Grad-CAM and BoVW-CAM methods.

Author's contribution

Arnaldo V. Barros da Silva: Writing, Investigation, Data collection and annotation, Experimental execution, Conceptualization.

Luis F. Alves Pereira: Supervision, Writing - Review & Editing, Guidance in methodology and experimental design, Conceptual advice.

Both authors contributed to the final manuscript, and all have read and approved the submitted version.

Competing Interests

The authors declare no competing interests relevant to the content of this article.

References

- Ahmed, M., Seraj, R., and Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295. DOI: 10.3390/electronics9081295.
- Akti, Ş., Tataroğlu, G. A., and Ekenel, H. K. (2019). Vision-based fight detection from surveillance cameras. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE. DOI: 10.48550/arXiv.2002.04355.
- Antipov, G., Berrani, S. A., Ruchaud, N., and Dugelay, J.-L. (2015). Learned vs. hand-crafted features for pedestrian gender recognition. DOI: 10.1145/2733373.2806332.
- Arroyo, R., Yebes, J. J., Bergasa, L. M., Daza, I. G., and Almazán, J. (2015). Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert systems with Applications*, 42(21):7991–8005. DOI: 10.1016/j.eswa.2015.06.016.
- Blunsden, S. and Fisher, R. (2010). The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(1-12):4. Available at: <https://homepages.inf.ed.ac.uk/rbf/PAPERS/unfbehavedata.pdf>.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. DOI: 10.1023/A:1010933404324.
- Chen, J.-H., Tseng, T.-H., Lai, C.-L., and Hsieh, S.-T. (2012). An intelligent virtual fence security system for the detection of people invading. *9th International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic and Trusted Computing*, pages 786–791. DOI: 10.1109/UIC-ATC.2012.64.
- Chen, M.-y. and Hauptmann, A. (2009). Mosift: Recognizing human actions in surveillance videos. DOI: 10.1184/R1/6607523.v1.
- Cheng, M., Cai, K., and Li, M. (2020). Rwf-2000: An open large scale video database for violence detection. DOI: 10.1109/ICPR48806.2021.9412502.
- Cian, D., van Gemert, J., and Lengyel, A. (2020). Evaluating the performance of the lime and grad-cam explanation methods on a lego multi-label image classification task. *arXiv preprint arXiv:2008.01584*. DOI: 10.48550/arXiv.2008.01584.
- Coşar, S., Donatiello, G., Bogorny, V., Garate, C., Alvares, L. O., and Brémond, F. (2016). Toward abnormal trajectory and event detection in video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):683–695. DOI: 10.1109/TCSVT.2016.2589859.
- Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague. Available at: <https://www.cs.cmu.edu/~efros/courses/AP06/Papers/csurka-eccv-04.pdf>.
- da Silva, A. V. B. and Alves Pereira, L. F. (2022). Bovwcam: Visual explanation from bag of visual words. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part II*, pages 45–55. Springer. DOI: 10.1007/978-3-031-21689-3_4.
- Delgado, B., Tahboub, K., and Delp, E. J. (2014). Automatic detection of abnormal human events on train platforms. *IEEE National Aerospace and Electronics Conference*, pages 169–173. DOI: 10.1109/NAECON.2014.7045797.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *IEEE conference on computer vision and pattern recognition*, pages 248–255. DOI: 10.1109/CVPR.2009.5206848.
- Deniz, O., Serrano Gracia, I., Bueno, G., and Kim, T.-T. (2014). Fast violence detection in video. volume 2. Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7294968>.
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302. DOI: 10.2307/1932409.
- Gao, Y., Liu, H., Sun, X., Wang, C., and Liu, Y. (2016). Violence detection using oriented violent flows. *Image and vision computing*, 48:37–41. DOI: 10.1016/j.imavis.2016.01.006.
- Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*. DOI: 10.48550/arXiv.1704.06857.
- Grego, M., Matiolański, A., Guzik, P., and Leszczuk, M. (2016). Automated detection of firearms and knives in a cctv image. *Sensors*, 16(1):47. DOI: 10.3390/s16010047.
- Hachiuma, R., Sato, F., and Sekii, T. (2023). Unified keypoint-based action recognition framework via structured keypoint pooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22962–22971. DOI: 10.48550/arXiv.2303.15270.
- Hao, W. and Zhili, S. (2020). Improved mosaic: Algorithms for more complex images. In *Journal of Physics: Conference Series*, volume 1684, page 012094. IOP Publishing. DOI: 10.1088/1742-6596/1684/1/012094.
- Hassner, T., Itcher, Y., and Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. *IEEE Computer Society Conference on Computer Vision*

- and Pattern Recognition Workshops (CVPRW), pages 1–6. DOI: 10.1109/CVPRW.2012.6239348.
- Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28. DOI: 10.1109/5254.708428.
- Jocher, G. (2020). YOLOv5 by Ultralytics. Available at: <https://github.com/ultralytics/yolov5>.
- Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics YOLO. Available at: <https://github.com/ultralytics/ultralytics>.
- Kang, M.-S., Park, R.-H., and Park, H.-M. (2021). Efficient spatio-temporal modeling methods for real-time violence recognition. *IEEE Access*, 9:76270–76285. DOI: 10.1109/ACCESS.2021.3083273.
- Keçeli, A. and Kaya, A. (2017). Violent activity detection with transfer learning method. *Electronics Letters*, 53(15):1047–1048. DOI: 10.1049/el.2017.0970.
- Khin, P. P. and Htaik, N. M. (2024). Gun detection: A comparative study of retinanet, efficientdet and yolov8 on custom dataset. In *2024 IEEE Conference on Computer Applications (ICCA)*, pages 1–7. DOI: 10.1109/ICCA62361.2024.10532867.
- Krausz, B. and Bauckhage, C. (2012). Loveparade 2010: Automatic video analysis of a crowd disaster. *Computer Vision and Image Understanding*, 116(3):307–319. DOI: 10.1016/j.cviu.2011.08.006.
- Lim, J., Al Jobayer, M. I., Baskaran, V. M., Lim, J. M., See, J., and Wong, K. (2021). Deep multi-level feature pyramids: Application for non-canonical firearm detection in video surveillance. *Engineering applications of artificial intelligence*, 97:104094. DOI: 10.1016/j.engappai.2020.104094.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer. DOI: 10.48550/arXiv.1405.0312.
- Lindeberg, T. (2012). *Scale Invariant Feature Transform*, volume 7. DOI: 10.4249/scholarpedia.10491.
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57. DOI: 10.48550/arXiv.1606.03490.
- Lu, D. and Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870. DOI: 10.1080/01431160600746456.
- Lu, N., Wu, Y., Feng, L., and Song, J. (2018). Deep learning for fall detection: Three-dimensional cnn combined with lstm on video kinematic data. *IEEE journal of biomedical and health informatics*, 23(1):314–323. DOI: 10.1109/JBHI.2018.2808281.
- Mohammadi, H. and Nazerfard, E. (2023). Video violence recognition and localization using a semi-supervised hard attention model. *Expert Systems with Applications*, 212:118791. DOI: 10.1016/j.eswa.2022.118791.
- Nanni, L., Ghidoni, S., and Brahnam, S. (2017). Handcrafted vs non-handcrafted features for computer vision classification. *Pattern Recognition*, 71. DOI: 10.1016/j.patcog.2017.05.025.
- Nievas, E. B., Suarez, O. D., García, G. B., and Sukthankar, R. (2011). Violence detection in video using computer vision techniques. *International conference on Computer analysis of images and patterns*, pages 332–339. DOI: 10.1007/978-3-642-23678-5_39.
- Olmos, R., Tabik, S., and Herrera, F. (2018). Automatic handgun detection alarm in videos using deep learning. *Neurocomputing*, 275:66–72. DOI: 10.1016/j.neucom.2017.05.012.
- Perez, M., Kot, A. C., and Rocha, A. (2019). Detection of real-world fights in surveillance videos. In *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2662–2666. IEEE. DOI: 10.1109/ICASSP.2019.8683676.
- Qi, D., Tan, W., Liu, Z., Yao, Q., and Liu, J. (2021). A gun detection dataset and searching for embedded device solutions. *CoRR*. DOI: 10.48550/arXiv.2105.01058.
- Rastogi, R. and Varshney, Y. (2024). *A comprehensive study for weapon detection technologies for surveillance using different YoloV8 models on primary data*, pages 241–268. De Gruyter, Berlin, Boston. DOI: 10.1515/9783111331133-013.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788. DOI: 10.48550/arXiv.1506.02640.
- Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2011). Robust video surveillance for fall detection based on human shape deformation. *IEEE Transactions on circuits and systems for video Technology*, 21(5):611–622. DOI: 10.1109/TCSVT.2011.2129370.
- Saba, T. (2021). Computer vision for microscopic skin cancer diagnosis using handcrafted and non-handcrafted features. *Microscopy Research and Technique*, 84:1272–1283. DOI: 10.1002/jemt.23686.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626. DOI: 10.1109/ICCV.2017.74.
- Setti, F., Conigliaro, D., Rota, P., Bassetti, C., Conci, N., Sebe, N., and Cristani, M. (2017). The s-hock dataset: A new benchmark for spectator crowd analysis. *Computer Vision and Image Understanding*, 159:47–58. DOI: 10.1016/j.cviu.2017.01.003.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. DOI: 10.48550/arXiv.1409.1556.
- Soliman, M. M., Kamal, M. H., El-Massih Nashed, M. A., Mostafa, Y. M., Chawky, B. S., and Khattab, D. (2019a). Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelli-*

- gent Computing and Information Systems (ICICIS), pages 80–85. DOI: 10.1109/ICICIS46948.2019.9014714.
- Soliman, M. M., Kamal, M. H., Nashed, M. A. E.-M., Mostafa, Y. M., Chawky, B. S., and Khattab, D. (2019b). Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85. IEEE. DOI: 10.1109/ICICIS46948.2019.9014714.
- Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*. DOI: 10.48550/arXiv.1212.0402.
- Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488. DOI: 10.1109/CVPR.2018.00678.
- Traoré, A. and Akhloufi, M. A. (2020). Violence detection in videos using deep recurrent and convolutional neural networks. In *2020 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 154–159. IEEE. DOI: 10.1109/SMC42975.2020.9282971.
- Van Gool, L. (2008). Action snippets: How many frames does human action recognition require? *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. DOI: 10.1109/CVPR.2008.4587730.
- Varam, D., Mitra, R., Mkadmi, M., Riyas, R., Abuhani, D. A., Dhou, S., and Alzaatreh, A. (2023). Wireless capsule endoscopy image classification: An explainable ai approach. *IEEE Access*. DOI: 10.1109/ACCESS.2023.3319068.
- Vijeikis, R., Raudonis, V., and Dervinis, G. (2022). Efficient violence detection in surveillance. *Sensors*, 22(6):2216. DOI: 10.3390/s22062216.
- Wei, K., Chen, B., Zhang, J., Fan, S., Wu, K., Liu, G., and Chen, D. (2022). Explainable deep learning study for leaf disease classification. *Agronomy*, 12(5):1035. DOI: 10.3390/agronomy12051035.
- Zhang, X., Shu, X., and He, Z. (2019). Crowd panic state detection using entropy of the distribution of enthalpy. *Physica A: Statistical Mechanics and its Applications*, 525:935–945. DOI: 10.1016/j.physa.2019.04.033.
- Zhou, P., Ding, Q., Luo, H., and Hou, X. (2017). Violent interaction detection in video based on deep learning. *Journal of Physics: Conference Series*, 844:012044. DOI: 10.1088/1742-6596/844/1/012044.
- Zou, Z., Shi, Z., Guo, Y., and Ye, J. (2019). Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*. DOI: 10.1109/JPROC.2023.3238524.