XLIII CILAMCE
21–25, November, 2022
FOZ DO IGUAÇU – PR – BRAZIL

# Detection and classification of firearms applied to entertainment media

Junior G. Santos[1], Gustavo M. de Almeida[1], Flávio G. Pereira[1]

[1] *Programa de Engenharia de Controle e Automação, Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo (IFES)*
Junior.guedes.s2021@gmail.com, gmaia@ifes.edu.br, fgarcia@ifes.edu.br

**Abstract.** Entertainment media have evolved considerably over the years. With this, the immersion and reality in the production of films, series and electronic games has increased. By improving this virtual reality, there are side effects in the production of content aimed exclusively at the younger audience, which, depending on the content, usually has an excessive load of violence, especially content related to firearms. Although the state of the art already has a significant advance in object detection through deep learning algorithms, real-time weapons detection is still a challenge. The detection of two classes of firearms was introduced to this work: handgun and heavygun. A dataset containing 2,000 images was used and another 2,000 were collected from various entertainment media and later annotated. The YoloV5 algorithm was used in this research, since it is already a consolidated model among researchers in the area. The analysis of the result is based on the exposure time in which firearms were exposed in entertainment content.

**Keywords: :** firearm, yolov5, entertainment, game, violence.

## 1 Introduction

Accidents resulting from firearms have grown significantly, which has drawn public health attention in the United States [1]. In 2018, nearly 39,000 Americans died from firearm-related accidents [2], with another 70,000 non-fatal injuries attributed to firearms [3]. People aged 15 to 24 have the highest firearm homicide rates [2] and are directly sensitive to media influences that put them at risk by exposing gun-related content [4, 5].

Violent television shows became common soon after TV became popular in American homes and are common to watch today, for example Gunsmoke, CSI and Miami Vice. Not so recently, video games, cell phones and internet videos have become more accessible as a form of entertainment by children and teenagers, being exposed to electronic game content such as Grand Theft Auto, Counter Strike and Resident Evil [6].

According to social learning theory [7], people learn successful behavior strategies by experience or by observing others. This theory exposes that people observe and copy or even imitate the behavior of others, which is called a model. The model can be a real person or a fictional character. In the case of this research, it is important to note that the model is especially likely to occur when the potential outcomes for a behavior are dangerous. Identification with the model is an important factor, and the more the person identifies with the model, the more he internalizes and simulates this behavior, especially when the model is rewarded for performing some behavior.

In view of this, due to the exposure that people undergo when choosing the aforementioned media as a source of entertainment, this research intends to create an algorithm capable of analyzing the exposure time of content related to weapons, using a neural network model. artificial weapon already consolidated and train it with firearms. The results will be displayed in the form of tables by comparing the proportional time in which each video was processed.

## 2 Related Works

Geetha [8] has developed an automatic video weapons detection system that is suitable for closed circuit television (CCTV), using the YOLOv3 (You only look once) algorithm for the purpose of real-time video weapons detection.

Jerong [9] proposed a framework for object recognition in a scenario where images are in low resolution through collaborative learning of two artificial neural networks. The proposed image enhancement network attempts to improve the extremely low resolution image into sharper and more informative images using collaborative learning signals from object recognition.

Xiao et al. [10] created an advanced tool for video forensic technical analysis to assist in forensic investigations. An adaptive video enhancement algorithm based on contrast-limited adaptive histogram equalization was introduced to improve the quality of CCTV images for use in forensic analysis. To aid video-based forensics, deep learning was applied to detect and track objects.

Harsh [11] presented a weapons detector using convolutional neural networks based on SSD and Faster RCNN algorithms. The proposed implementation uses two types of datasets. One dataset already has tagged images and the other is a set of manually tagged images. The results are shown in tables, both of which reached a good accuracy, but the application of the algorithms in real situations can be chosen based on the need between speed and accuracy.

## 3 Proposed Framework

This research aims to use a fast algorithm for detecting and classifying objects through deep learning techniques. The YoloV5 family of algorithms, developed by Jocher et al. [12], in its "s" version. YOLO models detect based on just one stage, where they classify and detect in just a moment and return the point of interest found with its respective bounding box. Among the various versions of the family of this artificial neural network model, the YoloV5s model was chosen, since the size of the network is smaller than the others, ensuring faster detection at the expense of a relatively lower accuracy.

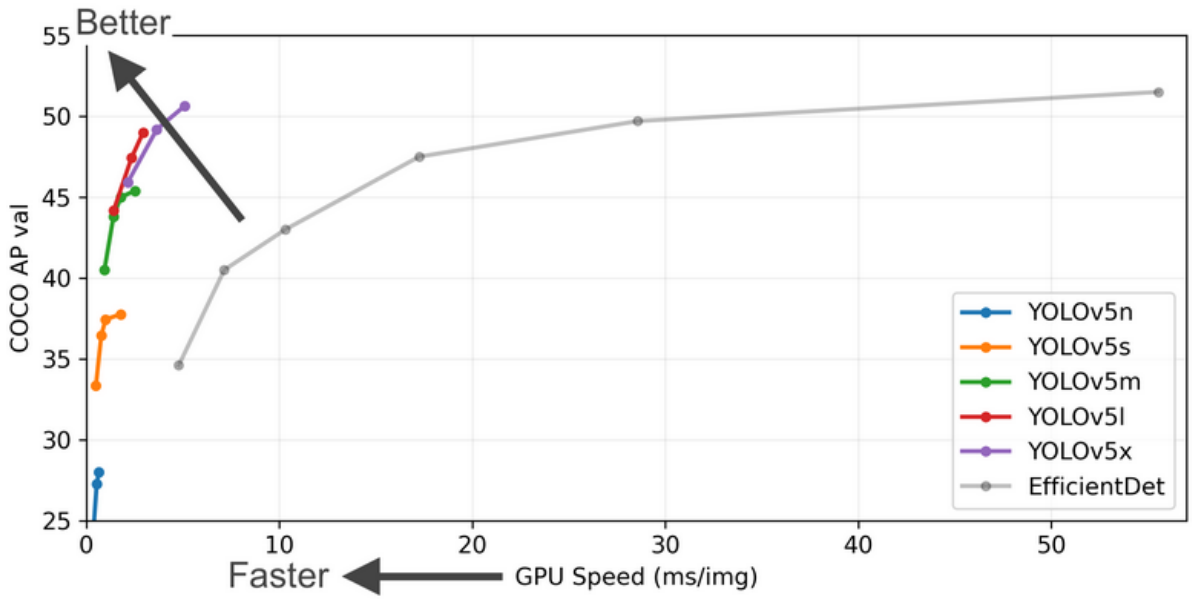


*Figure 1 Comparison between models YoloV5*

## 4 Data pre-processing

One of the most crucial steps in the training of an ANN is the organization and separation of the dataset, being one of the most time consuming steps. For this case, a set of 2,000 images were selected, in addition to the

availability of another 10% of the total set of images to be used as background images, with all images in the set being resized to the proportion of 256x256 pixels, ensuring fast training.

### 4.1 Model Architecture

The architecture of the YOLO v5 model works through single-stage detection and is subdivided into three important parts, being the backbone, neck and head.

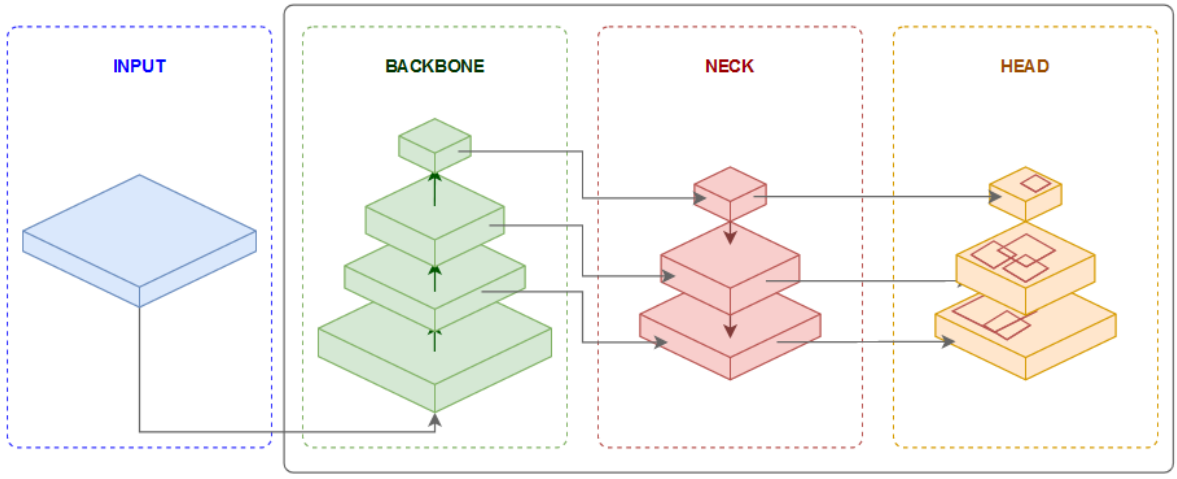


Figure 2. Object detection process

The backbone model exposed in Figure 2 is used to extract important features from the inserted image. The neck model is used to create the feature pyramid which is very useful so that the mobdel can generalize when the object changes size. In cases where data has not yet been seen, the feature pyramid helps the model to perform better. Finally, the head model is mainly used to perform the final detection. It is in this model that the final value is returned with the vector and the probabilities of classes, object scores and bounding boxes.

## 5 Experiments

### 5.1 Dataset

An open-source dataset made available by Olmos [13] was chosen, containing about 2,000 pistols annotated in the Pascal-VOC model, in addition to collecting images of pistols and revolvers on research sites and videos whose content does not have any copyright, being a total of 2,000 images of pistols, revolvers and heavy weapons collected. The manually annotated image dataset was manipulated and annotated in the RoboFlow tool, which is an online tool with a collection of modules that help organize, annotate, train and prepare the entire dataset. The open-source dataset was uploaded to Roboflow and later converted and exported in this tool to be read correctly in the format expected by YOLO v5, in addition to resizing all images to 256x256 resolution. No data augmentation technique was used for this experiment. After joining the two datasets into just one, it was subdivided into 70% for training, 20% for validation and 10% for testing. The total number of images available for this search is 7,000 images.

### 5.2 Platform

This research used the Google Collaboratory software which is a based product that allows users to execute code written in python directly through the browser. Although there is a free version, the PRO version was used for this research, which guaranteed a GPU with 15GB of memory and a little more speed in the training and tests performed.

Google Drive was used to store the dataset, since it has direct integrations with Colaboratory, in addition to guaranteeing sufficient capacity to store the files of this present work.

### 5.3 Model evaluation

This research uses precision (P) and recall (R) and mean precision (mAP) to evaluate the performance of objects detected by the model. The expressions for P and R are as follows:

$$P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN}$$

Precision measures how accurate the inferences are. Recall measures the number of times true positives are inferred. AP is an average accuracy rate, which is an area under the P-R curve. mAP is the average of AP, that is, it is the sum of each AP category, defined by the number of categories.

$$AP = \int_0^1 P(R)dR$$

$$mAP = \frac{1}{|Qr|} \sum_{q=Qr} AP(q)$$

Qr is the number of categories.

### 5.4 Training

Considering that the network will be able to detect and classify weapons in video games and in videos related to this type of interactive content, 3D modeled weapons images were also selected. The training dataset was composed as shown in Figure 3 below.

Figure 3. Dataset images

To train the YOLO network, the default hyperparameters of the model were defined, changing minimum training settings.

In the first training, 4,000 images were associated with the dataset, with the maximum size of each image being set to 256x256, batch size of 384, with 400 epochs. The training lasted 1h42min, obtaining a mAP@0.5 of 80%.

In the second training, 400 more images were added to be used as background, the size and batch size defined previously were maintained, but the number of epochs was changed to 700. The training lasted 2h17min, obtaining mAP@0.5 of 89%.

In all, 10% more random images were added to be used as backgrounds. Background images are images that do not have any objects that will be trained in a way that reduces false positives. No markup is used in this type of image, just include them in the training folder of the YOLO network. Figure 4 presents examples of used background images.

*Figure 4 Background images*

### 5.5 Results and evaluations

In this section, the results obtained from the different channels in which the model made the inferences will be explored. The videos were taken from popular channels on the Youtube portal. In order to be impartial in the analysis, a filter was defined containing all the main live games, considering the subject "game", in any country and in any language. The 10 most viewed videos on the portal were analyzed according to automatic classification. In addition, each video had only its first 12 minutes analyzed. At the end of the analysis of the 10 videos, the average of the total number of times in which weapons were detected was calculated. All selected videos have an average of 27 frames per second.

### 5.6 Method of analysis

The videos were delivered to the network to be analyzed so that the seconds in which the weapons were exposed over the time the video was viewed were counted. Time was counted in seconds and the frame rate per second was used to measure time.

The video is fragmented into frames and, as the classes are found by the network, then the number of frames in which an object was exposed is counted. At the end of each analysis, a simple calculation is performed to find the total seconds in which the classes were displayed according to the formula below:

$$S = \frac{t}{i}$$

Where t is the video frame rate per second and i is the total number of frames in which the classes were viewed at least once.

### 5.7 Evaluation

The two classes have each 2000 images. The evaluation results obtained were:

Table 1. Results by class

| Class | P | R | mAP | mAP95 |
|---|---|---|---|---|
| Handgun | 0.887 | 0.768 | 0.826 | 0.54 |
| Heavygun | 0.941 | 0.916 | 0.954 | 0.69 |

Table 2. Results by indicators

| Metric | Value |
|---|---|
| Best/epoch | Epoch 635 |
| mAP 0.5 | 0.887 |
| mAP 0.5:95 | 0.610 |
| Precision | 0.927 |
| Recall | 0.842 |

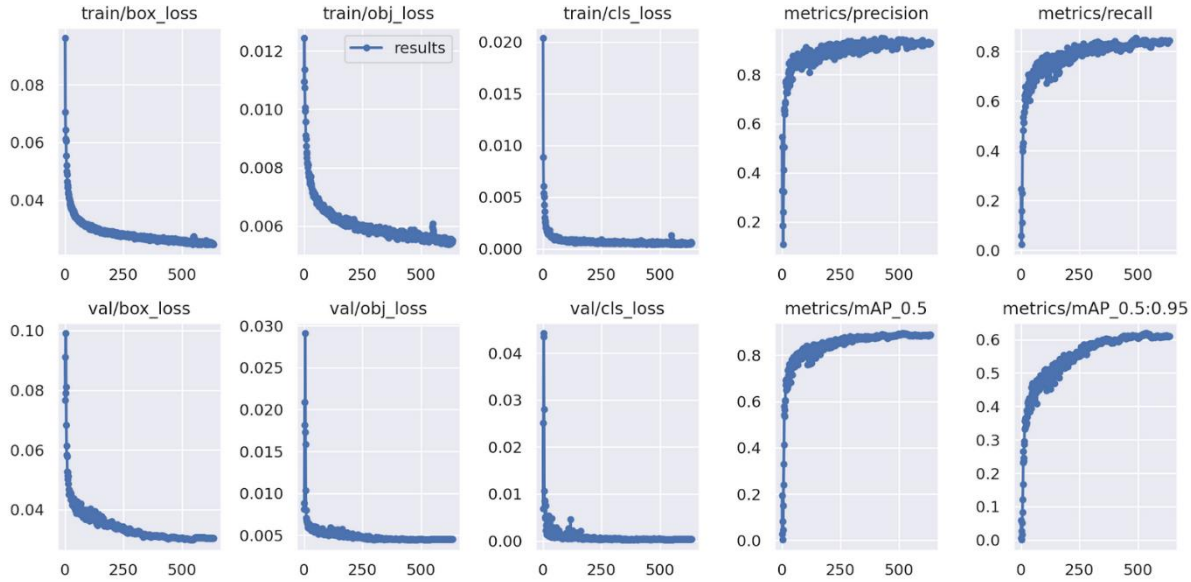


*Figure 5 Results in chart*

### 5.8 Videos Results

At the end of the validations, the videos were submitted to the trained model, obtaining the following results:

Table 3. Video results

| Title | FPS (t) | Time | Exposure (s) | % |
|---|---|---|---|---|
| Gang Shootout \| GTA V | 25 | 555s | 176s | 31% |
| O Assalto Ao Cayo Perico | 25 | 614s | 221s | 35% |
| Crise Explosiva – Movie \| GTA 5 Swat | 23 | 603s | 246s | 40% |
| Battlefield V - 4k MAX Ray Tracing Settings | 30 | 1244s | 308s | 24% |
| Diamond Assault Defenders | 30 | 624s | 185s | 29% |
| 911 calls, bodycam video released of deadly Phoenix police shooting | 30 | 83s | 15s | 18% |
| Body camera footage from Jackson County officer involved shooting | 30 | 236s | 17s | 7% |

## 6 Conclusions

Gun detection is widely used in public security for real-time monitoring of city neighborhoods. In this century, where the automation of various tasks is sought, the detection of weapons has become a very interesting field of research and showing good results. In the case of this work, we tried to achieve the same objective, however, focused on detecting and classifying weapons in entertainment media. The results were not satisfactory due to the small amount of images in the dataset. In video games, although it detected the weapons as expected, the model failed to detect it in some cases and, in other cases, resulting in a false positive.

In situations where the objective is to detect weapons aimed at public safety, it was observed that the dataset usually has images from cameras in which the angle changes very little. In entertainment content such as video games, newspapers and similar content, the image angle changes considerably, requiring a rich dataset from many different angles.

With this, it is concluded that it is feasible to develop an algorithm to detect this type of object, taking into account a more comprehensive dataset in order to improve accuracy and reduce false-positive cases. In addition, it is important to consider the inclusion of video game images that have a wide range of weapons, being possible

to modify them by the players themselves.
For future work, this algorithm can be used in order to contribute to other researchers in the humanities, being an auxiliary tool in research on human behavior and related areas.

**Authorship statement.** The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

## References

[1] Bauchner H, Rivara FP, Bonow RO, Bressler NM, Disis ML, Heckers S, et al. Death by gun violence—A public health crisis. JAMA Psychiatry. 2017; 74(12):1195–6.
[2] CDC. Fatal injury reports, national, regional and state, 1981–2018 [Internet]. Atlanta, GA: National Center for Injury Prevention and Control; 2018 Aug. Available from: https://webappa.cdc.gov/sasweb/ncipc/mortrate.html
[3] Everytown for gun safety. A more complete picture: The contours of gun injury in the United States [Internet]. 2019 p. 1–7. Available from: https://everytownresearch.org/a-more-complete-picture-the-contours-of-gun-injury-in-the-united-states
[4] Anderson CA, Bushman BJ. The effects of media violence on society. Science. 2002; 295:2377–9. https://doi.org/10.1126/science.1070765
[5] Bushman BJ, Huesmann LR. Short-term and long-term effects of violent media on aggression in children and adults. Arch Pediatr Adolesc Med. 2006; 160(4):348–52. https://doi.org/10.1001/archpedi. 160.4.348
[6] Huesmann LR. The impact of electronic media violence: scientific theory and research. J Adolesc Health. 2007 Dec;41(6 Suppl 1):S6-13. doi: 10.1016/j.jadohealth.2007.09.005.
[7] Bandura A. Social learning theory of aggression.J Commun. 1978;28(3):12-29. doi:10.1111/j.1460-2466.1978. tb01621.x
[8] Akash Kumar. K. S , Akshita. B. P , Arjun. M , Dr. N. Geetha, 2021, Weapon Detection in Surveillance System, international journal of engineering research & technology (IJERT) Volume 10, Issue 05 (May 2021)
[9] JEONGIN SEO AND HYEYOUNG PARK School of Computer Science and Engineering, Kyungpook National University, “Object Recognition in Very Low Resolution Images Using Deep Collaborative Learning”,IEEE Access (2020).
[10] J. Xiao, S. Li and Q. Xu, "Video-Based Evidence Analysis and Extraction in Digital Forensic Investigation," in IEEE Access, vol. 7, pp. 55432-55442, 2019
[11] H. Jain, A. Vikram, Mohana, A. Kashyap, Ayush Jain Telecommunication Engineering, “Weapon detection using AI and deep learning for security application”, IEEE (2019)
[12] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, et al. “ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference (v6.1)”. 2022, available from: https://doi.org/10.5281/zenodo.6222936
[13] Olmos, R., Tabik, S., & Herrera, F. (2018) Automatic handgun detection alarm in videos using deep learning. Neurocomputing, 275, 66-72. doi.org/10.1016/j.neucom.2017.05.012
[14] Yao, J.; Qi, J.; Zhang, J.; Shao, H.; Yang, J.; Li, X. A Real-time detection algorithm for kiwi fruit defects based on YOLOv5. 2021. Available from: https://www.mdpi.com/2079-9292/10/14/1711