Detection of weapon possession and fire in Public Safety surveillance cameras

Natan Santos Moura[1], João Medrado Gondim[1], Daniela Barreiro Claro[1]
Marlo Souza[1], Roberto de Cerqueira Figueiredo[1]

[1]FORMAS - Formalismos e Aplicações Semânticas
Instituto de Computação – Universidade Federal da Bahia (UFBA)
Campus de Ondina – 40170-110 – Salvador – Bahia – Brasil

{natan.moura,joao.gondim,dclaro,msouza1,roberedo}@ufba.br

Abstract. *The employment of video surveillance cameras by public safety agencies enables incident detection in monitored cities by using object detection for scene description, enhancing the protection to the general public. Object detection has its drawbacks, such as false positives. Our work aims to enhance object detection and image classification by employing IoU (Intersection over Union) to minimize the false positives and identify weapon holders or fire in a frame, adding more information to the scene.*

1. Introduction

Public safety involves protecting the general public against threats that could harm and endanger people [University 2019]. In Brazil, according to article 144 of the Brazilian Federal Constitution [BRASIL 2021], public safety is the responsibility of state agencies such as the federal police forces, state military, and civilian police forces as well as the military firefighters. The use of surveillance cameras by public safety agencies to protect the public has been growing. It can identify criminal incidents, respond to critical incidents, monitor pedestrian and vehicle traffic activity, document officer and offender conduct during interactions, and assist in identifying offenders [Lexipol 2021]. The city of Palotina, in Paraná, which has a surveillance camera for every 59 inhabitants, was considered one of the most monitored cities in Latin America and, as a result, had its crime rate reduced by 80% [Eletronica 2021].

Computer vision techniques can be employed to identify criminal and critical incidents in surveillance cameras for automatic detection of objects related to crimes and image classification. This work proposes composing different computer vision methods to object detection and image classification to public safety incidents, automatically detecting guns, knives, fire, and people. These methods are part of a research project called Cloud Security Interoperable Society (CSIS), conceived in partnership with public safety agencies (Military Police, Civil Police, and Fire Department).

The occurrence of false positives impacts computer vision for detecting objects such as guns and knives. The high rate of false positives makes detecting objects in surveillance cameras an open problem. Besides, automatic detection of objects related to crimes in surveillance cameras has some limitations. While detecting objects from images retrieved from surveillance cameras adds more information to the general understanding of the scene, it might not be enough for proper knowledge acquiring. For example, a handgun detected on a scene might be held by a person, on a police officer's holster, an advertisement on an outdoor, or even a false positive wrong detection. A single object detected does not characterize a criminal incident scene. The manipulation of objects by people is necessary to suppose a possible crime scene. Then, detection techniques must consider the union between the gun object and the person object to infer that the same manipulates a gun. The detection itself is important but can be more precise by using more than one inference to add to the scene. For instance, linking a detected weapon to a detected person on the frame can be done by analyzing the intersection of both objects detected.

Our goal is to integrate and facilitate the detection of public safety incidents through a pipeline to detect a weapon possession and the presence of fire in a scene, addressing the message to their respective agencies: Military Police or Fire Department, respectively.

This article is organized as follows. Section 2 describes the image classification and object detection background. Section 3 presents our models and results. Section 4 discusses our findings, and in section 5, we present some ethical issues, finalizing with section 6.

2. Background

In this section, we describe each of the methods for inferring and gathering information from the video surveillance cameras to ease the reading of the whole paper since we use different computer vision methods in the pipeline.

2.1. Image classification

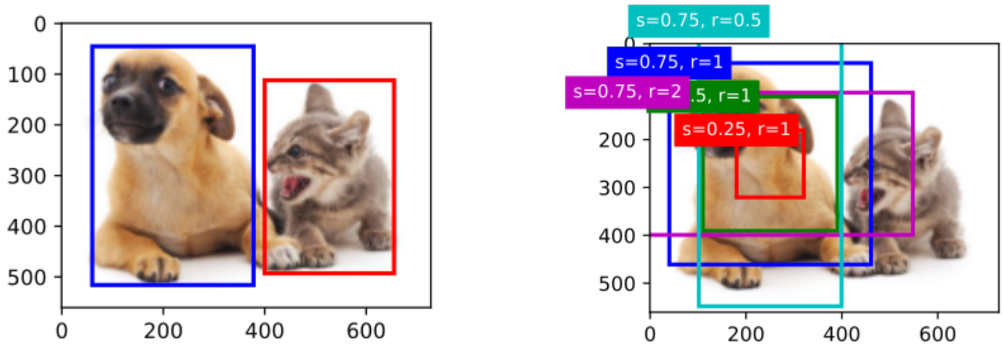
Image classification is the process of categorizing all the pixels in an image to obtain a set of labels [Al-Doski et al. 2013]. Systems performing this task commonly employ two relevant tools: the feature extractors and the classifiers.

Feature extraction consists of extracting relevant features from an image and define its label [Medjahed 2015]. It is possible to classify an image based on its content, thus reducing the amount of analyzed data, extracting only relevant properties, organizing and defining classes to the content, allowing classifiers/detectors to recognize if one frame from a video has or not a gun in it, for instance. There are many feature extraction techniques based on color, texture, shape, etc. A feature extractor can process an image as input then produce labels as output, for instance, the presence of a knife.

Image classification depends not only on feature extraction but also on other classifications tasks. The classification process consists of comparing predefined samples (usually from a dataset) to a pattern extracted from a different image [Rama Gaur 2017]. Examples of classification methods are: Decision Tree, Support Vector Machine (SVM), and Neural Networks [Dey et al. 2014].

2.2. Object detection

One of the most important concepts for object detection is the bounding box. A bounding box is a rectangle determined by x (upper-left corner) and y (lower-right corner) coordinates, followed by width and height to define the spatial location of an object. We can



(a) Examples of bounding boxes [Zhang et al. 2021]

(b) Examples of anchors [Zhang et al. 2021]

understand object detection as defining the position (through bounding boxes) and the label of objects in a given image. Different from image classification, there are often multiple dispersed objects, and the object detection tools can detect and set a label for each one [Zhang et al. 2021].

The anchor boxes are the central concept to understanding the object detection models. Anchor boxes can be defined as bounding boxes with varying scales and aspect ratios centered on each pixel of a training image. They are essential to determine regions on the training image, the edges of these regions predict the correct predetermined object bounding box. The *Intersection over Union* metric is commonly employed to evaluate the anchor box regions' accuracy in predicting an object[Zhang et al. 2021].

Intersection Over Union (IoU) is one of the most employed metric for object detection [Rezatofighi et al. 2019]. It defines the "correct detection" [Padilla et al. 2020]. This technique is based on the area of the bounding box and their relation. The bounding box can be a "ground-truth box", previously annotated object, or a "prediction box" predicted by the object detection algorithm. The IoU is computed by the union of the overlapped area between the ground-truth and the prediction box [Padilla et al. 2020].

The IoU value is compared to a threshold to be correct (True Positive - TP) or incorrect (False Positive - FP). The value of the threshold could be defined based on the needed level of accuracy.

2.3. YOLO

YOLO (You Only Look Once) is a real-time object detection system released in 2016 by Joseph Redmon. The goal of YOLO is to recognize objects faster than usual convolutional neural networks, without losing accuracy, by looking only once at the image, setting the object detection as a single problem. The pipeline resizes the input, runs it in a single convolutional neural network, and thresholding the results by model's confidence [Redmon et al. 2016]. YOLO divides the image into multiple sub-regions and sets five anchor boxes to each one to perform the detection. The probability of a specific object is calculated, and the region with the highest probability is selected [Kanehisa and Neto 2019].

3. Related work

The detection of guns and knives in surveillance videos has been a challenge. As far as we know, no authors have employed a combination of object detection models to provide gun possession information. Our main contribution in this work is to combine knives and guns detection with persons bounding boxes to identify weapon possessions.

3.1. Weapon detection model improvement

Some related works focus on finding techniques to enhance the accuracy when detecting weapons in images.

Authors in [Olmos et al. 2017] explored two approaches for automatic handgun detection: sliding window with Histogram of Gradients and region proposal (with Faster R-CNN). Using a newly created dataset (Pistol Detection), they trained both models. They defined a metric (Alarm Activation Time per Interval) for assessing detection models used for automatic detection systems in videos.

The work in [Fernández-Carrobles et al. 2019] employed Olmos dataset to explore different architectures using Faster R-CNN as the primary object detection tool. SqueezeNet obtained better results than Olmos' VGG-16 based on the deep learning model when detecting guns. As for knife detection, they showed some improvements when using GoogleNet as architecture for Faster R-CNN.

Authors from [Gelana and Yadav 2019] employed a background subtraction and edge detection to train a CNN (Convolutional Neural Network) classifier and applied it with the sliding window technique. Although reporting good results, the training and testing dataset is built from good quality CCTV videos, which is not always feasible, limiting the use of the algorithm.

To improve the images to detect handguns, authors in [Ruiz-Santaquiteria et al. 2020] came up with a novel method to train YOLOv3 with binary images using body pose key points extracted using OpenPose, a multi-person pose estimator. This kind of training aims to feed the neural network with the weapon's appearance and the person's body pose.

3.2. Dataset improvement for object detection

These related work concerns datasets to improve inference on images for video surveillance cameras.

Authors in [Pérez et al. 2020] explored the problem of false-positive detections due to small objects handed similarly to guns or knives (such as wallets, smartphones, or credit cards) on images using model ensembling with binarization techniques (One Versus One and One Versus All) along with an object detection model (Faster R-CNN) and the creation of a dataset-specific for Small Objects Handed Similarly to a Weapon (SOHAS Weapon dataset). Using object detection and image classification, they achieved good results compared to a baseline detector based on Faster R-CNN on avoiding False Positive alarms. Similar to this work, we also detect objects, and we do image classification.

Authors in [González et al. 2020] analyzed the problem of detecting small weapons due to their representation of few pixels and having few annotations on datasets. This is an important difference from the previous work because CCTV cameras have lower image resolution; thus small objects are represented by fewer pixels, which impairs the task. Using a strategy of two stages training and a newly dataset simulating a fake attack (Mock Attack Dataset) to annotate images of weapons on video surveillance cameras, they also showed the impact of computer generated dataset on training models for weapon detection by creating a Unity [1] engine based dataset and using a synthetic dataset built by Edgecase.ai[2].

Authors in [Lim et al. 2021] provide a new dataset (Monash Guns Dataset), taking into consideration some design tips from ImageNet dataset and MS COCO and aiming to mix images of weapons in canonical and non-canonical situations. After the creation of the dataset, they trained a multi-level multi-scale object detector implemented by M2Det.

3.3. Fire image classification

Gathering images to analyze better information on emergencies [Cazzolato et al. 2017] came up with FiSmo-Images (Fire and Smoke Images), which is a dataset composed of four others with images collected from Flickr, YouTube, and simulations.

Authors from [Muhammad et al. 2018] explored different CNN architectures to study the trade-off between accuracy and computational cost when classifying whether an image has a fire on it or not. Their work mainly focused on video surveillance systems, giving attention to low computational cost models with good accuracy.

4. Intersection of Objects in CSIS

Our work proposes to send messages to a center of alerts forwarding from images collected from surveillance video cameras (CCTV cameras) to describe the scenarios such as models for:

- Image classification for classifying images with/without fire;
- Object Detection for detecting weapons focusing on pistols/handguns and knives for now and people in the images. Such detections can be used to describe *who* is holding a gun or even to minimize false positives in images on cases where a gun or knife is detected where there is no person near it.

We intend to use a combination of models to get these information: YOLOv5 [Jocher et al. 2021] (PyTorch backend) for object detections (knives and handguns) and Tensorflow [Rosebrock 2019] for image classification (fire or no fire).

4.1. Models

We followed the architecture proposed in [Rosebrock 2019] to identify the presence of fire in images. The authors employed SeparableConv2D (separable convolution) for having good inference performance with little computational power. Instead of SeparableConv2D, we employed Conv2D (default convolutional layer). The hyperparameters were the same as the authors.

For the task of object detection, we employed YOLOv5[3], since it has different architectures depending on some parameters of the model. Detecting such small objects as handguns or knives is a challenging task, which YOLOv5x was chosen for training

[1] https://unity.com/

[2] https://www.edgecase.ai/

[3] https://github.com/ultralytics/yolov5

and inference. The hyperparameters default from YOLO was employed but with a batch size of 8 images and just 100 epochs. Such configurations were due to a limitation on the Google Colab environment where we performed our empirical validation.

YOLOv5x was trained with COCO (Common Objects in Context) [Lin et al. 2015] for detection of the classes person/car/knife as these classes are already present on COCO. No specific CCTV images were trained. All weights were downloaded from YOLOv5 repository.

4.2. Datasets

Considering the problem of detecting guns in video surveillance cameras, we intended to use a training dataset as close as possible to images found on images from security videos. The Monash Guns Dataset [Lim et al. 2021] has weapons in a representative way and also with a variety of different guns. Authors describe the dataset as a mixture of canonical and non-canonical images of guns recorded similar to video surveillance cameras. Due to these characteristics, we applied this dataset.

For the image classification task, between "fire" and "non fire" frames, we enlarged the dataset from [Cazzolato et al. 2017] with images of CCTV found on a Kaggle [4] fire detection on surveillance cameras competition. The idea is to have a representative dataset of fire images in the wild and video surveillance cameras recording fires. When the training happens, our final model can correctly classify images containing a fire in both situations. Images for the knife detection task are in the COCO dataset, since we employed the default weights from YOLOv5.

4.3. Package compositions

The pipeline of machine learning models intends to identify, detect and send messages about different threatening situations. For this, the application receives frames generated by a video camera and applies each trained model to them. Information retrieved by the model to provide alerts to the CSIS dashboard, Figure 2 describes this process.

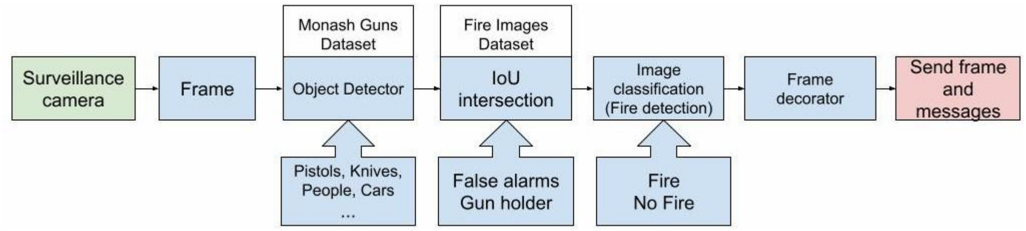


Figure 2. Process pipeline.

Inferences made by the object detector return a class and the bounding boxes of each object. Therefore, we have more information on what may be happening on a frame and enable the usage of such information to solve different tasks on the scene.

We can handle the bounding box of one person and the bounding box of a gun together. Within this, we identify the weapon holder by searching which person's bounding

[4] https://www.kaggle.com/ritupande/fire-detection-from-cctv

box has the greatest IoU overall. The use of bounding boxes can minimize some false detections, such as a weapon with an IoU value of 0 with a person detected in a frame.

The output from the fire image classification model is binary and easy to use, either the image has a fire on it or not. However, using some characteristics from neural network classifiers, we might be able to know more. François Chollet [F. 2017] explains how to use the outputs of a convolutional layer to determine what part of an image activated this layer on a given class, thus been able to ”see” what triggered our fire classifier in case of fire detection.

Each model on the sequence retrieves some information from the frame. After gathering all of these data, we use them to describe the scene better to send helpful information for CSIS. The use of ensemble models allows different information gathering from a single image.

5. Results

Some samples are provided to describe our results. Figure 3 and Figure 4 show the IoU information comparing the rectangle of weapon with every bounding box of a person detected in the scene thus, associating the weapon with the person carrying it.

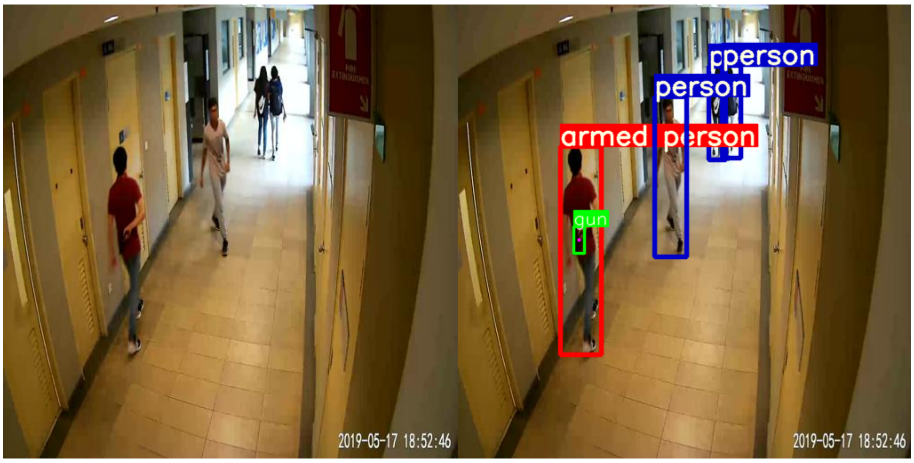


Figure 3. Detecting person holding gun using IoU.

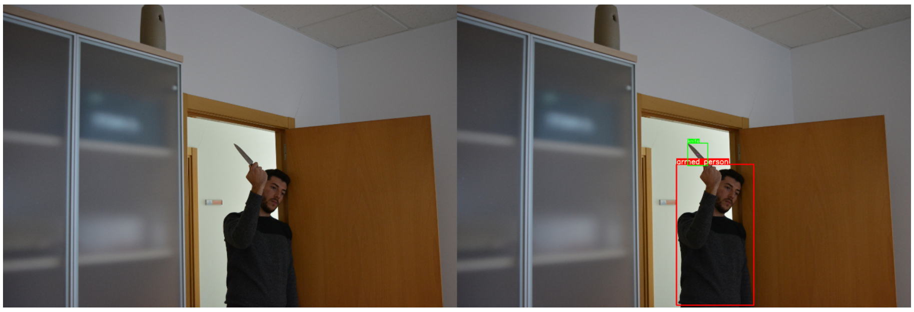


Figure 4. Knife detected and linked to the person holding it.

Figure 5 misdescribes a car detected as a weapon. It has an IoU value of 0 with the person on the scene; we can then classify it as a false positive.

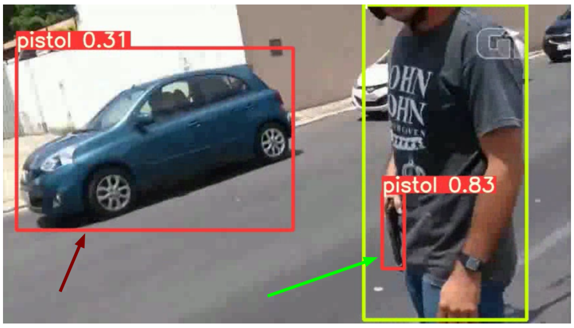


Figure 5. False alarm of gun detected due to no IoU with a person and a gun inside a person's bounding box.

Figure 6 depicts the output of the fire classification model to determine which part of the image activates the classifier and the combination of such with YOLOv5's output to determine that the car is on fire as the red part of the heatmap is inside the truck's bounding box.

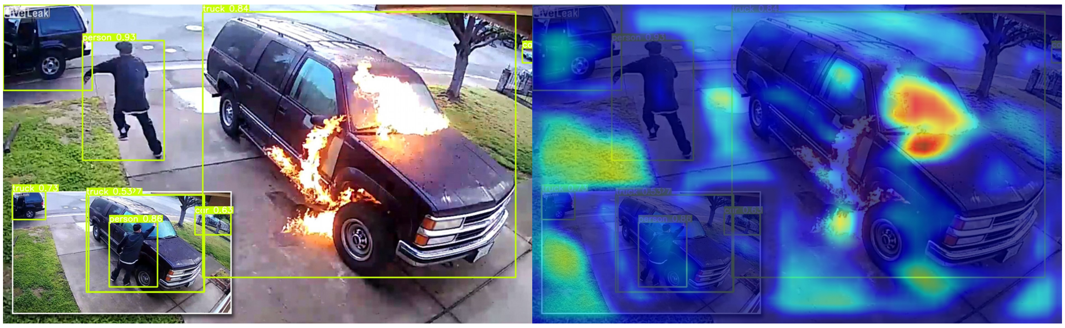


Figure 6. Image of a car on fire, followed by YOLO bounding boxes predictions and a heatmap made from our fire classification image neural network.

6. Discussion

The combination of methods increases the possibility of describing the scene, leading to a more robust detection of incidents. This is important when dealing with applications in the public security domain as CSIS. However, the way this information is used must be well analyzed since restricting their use might lead to misunderstandings about what is happening in the frame. As an example, even though giving us important information, applying IoU for understanding scenes might lead to wrong descriptions as well and, therefore, must be used keeping in mind some aspects: if the person holding the weapon on the scene raises their arm beyond their bounding box limits, the IoU between them and a gun might be equal to 0, wrongly indicating that this person is not holding a gun or even turning a right detection into a false negative if we determine that guns with 0 IoU with a person's bounding box indicate a wrong prediction as shown in Figure 7.

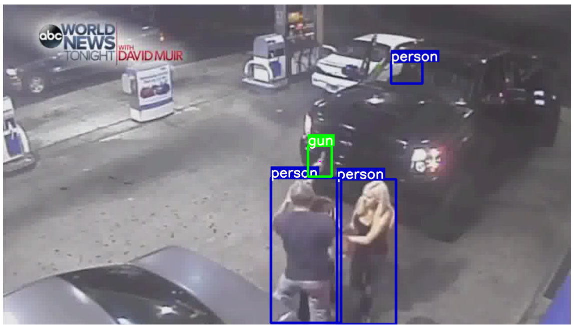


Figure 7. Mistake on IoU use due to the gun holder's arm being raised.

Another issue is the proximity of bounding boxes for different people detected in the same picture. The value of IoU is employed to determine who is holding a weapon. If two people are close together (even if we count the perspective of the camera), the one not carrying the gun might be classified as "armed person" due to the intersection of bounding boxes, Figure 8 exemplifies this.

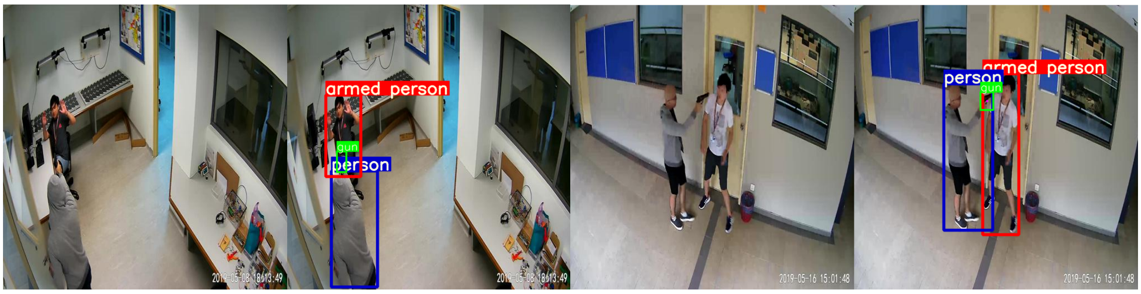


Figure 8. False detection of person holding gun due to the proximity of bounding boxes.

These questions can be dealt with different approaches depending on what we are trying to mitigate and what are our resources. In case we can get body pose keypoints from the image, even if the IoU suggests a wrong classification of the holder and the gun by calculating the distance between the detected gun and the points of hands, we might be able to correct the mistake as shown in Figure 9. Here the body pose keypoints were obtained with OpenPifPaf [Kreiss et al. 2021].

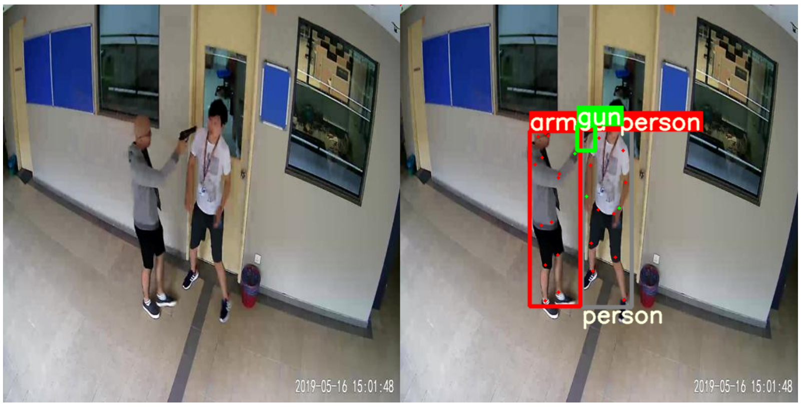


Figure 9. Correct detection of the person possessing the gun.

To be able to get more details on the frame, some ideas came on the making of the project, such as using the keypoints obtained to train an SVM classifier into ”holding gun” or ”unarmed”, this would help to avoid misclassifications on which person is holding an weapon; using the keypoints to detect where the person holding the gun is pointing it to, identifying is someone is in imminent danger; train some image classifier on a dataset like DeepFashion[5] and try to identify what the person holding a gun is wearing to be able to better describe them to the security force being called.

7. Ethics aspects analysis

Some ethical issues related to discrimination and privacy violation arise in situations involving the use of surveillance technology and intelligent systems in public safety.

Authors in [Wilson et al. 2019] analyzed the tone of the pedestrian skin as decisive for the performance of a pedestrian detection algorithm. An object detector on surveillance cameras can generate a false positive for a gun if a black person is not handling a gun. Thus, we can consider that our object detector in surveillance cameras needs to analyze skin tone. To mitigate this discriminatory bias, it is necessary to incorporate actions that evaluate hypotheses and find evidence of the bias in your project. Authors in [Rovatsos et al. 2019] used statistical approaches that focus on identifying patterns of discrimination in datasets before the training stage or adjusting the model’s outputs. Authors in [Raji et al. 2020] used a benchmark dataset to audit intelligent facial recognition algorithms.

Surveillance cameras can be invasive and violate people’s privacy. The internal environment of a residence can be equipped with several surveillance cameras. Cameras can also be placed in public buildings. Providing such images to public safety agencies in their command centers and patrol cars via the internet increases the potential that more people will see images. This may constitute a violation of people’s privacy [Slobogin 2003]. To mitigate this privacy violation, conduct rules in surveillance and sanctions imposed by their violation should be publicized in laws [Slobogin 2003].

8. Conclusion and Future Work

This work is part of a Public Security project called CSIS, which aims to assist the security forces in detecting objects with a human possession. Our findings encourage the direction by using a composition of models to detect together two or more approaches. As future work, we envision providing an *ensemble* method to evaluate it against single ones.

Acknowledgement

We would like to thank CNPQ financial support (PIBIC-AF 2020/2021), FAPESB financial support (BOL and TIC013/2015), and CAPES Financial Support Code 001.

References

Al-Doski, J., Mansorl, S., and Mohd, H. (2013). Image classification in remote sensing. *J. Environ. Earth Sci.*, 3(10):141–148.

[5] http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html

BRASIL (2021). Constituição da república federativa do brasil de 1988. https://www.senado.leg.br/atividade/const/con1988/CON1988_05.10.1988/art_144_.asp. [Online; Last accessed 10 Aug 2021].

Cazzolato, M., Avalhais, L., Chino, D., Ramos, J., Souza, J., Rodrigues Jr, J., and Taina, A. (2017). Fismo: A compilation of datasets from emergency situations for fire and smoke analysis.

Dey, N., Mishra, G., Kar, J., Chakraborty, S., and Nath, S. (2014). A survey of image classification methods and techniques.

Eletronica, R. (2021). Cidade mais monitorada do brasil, palotina reduz taxa de criminalidade em 80%. https://revistasegurancaeletronica.com.br. [Online; Last accessed 10 Aug 2021].

F., C. (2017). *Deep Learning with Python*. Manning Publications Co.

Fernández-Carrobles, M., Deniz, O., and Maroto, F. (2019). Gun and knife detection based on faster r-cnn for video surveillance.

Gelana, F. and Yadav, A. (2019). Firearm detection from surveillance cameras using image processing and machine learning techniques: Proceedings of icsiccs-2018.

González, J. L., Zaccaro, C., Alvarez-Garcia, J., Soria Morillo, L., and Caparrini, F. (2020). Real-time gun detection in cctv: An open problem. *Neural networks : the official journal of the International Neural Network Society*, 132:297–308.

Jocher, G., Stoken, A., Borovec, J., NanoCode012, Chaurasia, A., TaoXie, Changyu, L., V, A., Laughing, tkianai, yxNONG, Hogan, A., lorenzomammana, AlexWang1900, Hajek, J., Diaconu, L., Marc, Kwon, Y., oleg, wanghaoyang0106, Defretin, Y., Lohia, A., ml5ah, Milanko, B., Fineran, B., Khromov, D., Yiwei, D., Doug, Durgesh, and Ingham, F. (2021). ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations.

Kanehisa, R. and Neto, A. (2019). Firearm detection using convolutional neural networks.

Kreiss, S., Bertoni, L., and Alahi, A. (2021). Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *CoRR*, abs/2103.02440.

Lexipol (2021). Public safety policy manual 2020. https://sunnyvale.ca.gov/civicax/filebank/blobdload.aspx?BlobID=26744. [Online; Last accessed 10 Aug 2021].

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.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

Medjahed, S. A. (2015). A comparative study of feature extraction methods in images classification. *International Journal of Image, Graphics and Signal Processing*, 7:16–23.

Muhammad, K., Ahmad, J., Lv, Z., Bellavista, P., Yang, P., and Baik, S. (2018). Efficient deep cnn-based fire detection and localization in video surveillance applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, PP.

Olmos, R., Tabik, S., and Herrera, F. (2017). Automatic handgun detection alarm in videos using deep learning.

Padilla, R., Netto, S., and da Silva, E. (2020). A survey on performance metrics for object-detection algorithms.

Pérez, F., Tabik, S., Castillo Lamas, A., Olmos, R., Fujita, H., and Herrera, F. (2020). Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance. *Knowledge-Based Systems*, 194:105590.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In Markham, A. N., Powles, J., Walsh, T., and Washington, A. L., editors, *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, pages 145–151. ACM.

Rama Gaur, D. V. S. C. (2017). Classifiers in image processing. *International Journal on Future Revolution in Computer Science and Communication Engineering*, 3:22–24.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression.

Rosebrock, A. (2019). Fire and smoke detection with Keras and Deep Learning. `https://www.pyimagesearch.com/2019/11/18/fire-and-smoke-detection-with-keras-and-deep-learning/`. [Online; Last accessed 10 Jun 2021].

Rovatsos, M., Mittelstadt, B., and Koene, A. (2019). *Landscape Summary: Bias in Algorithmic Decision-Making: What is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?* UK Government.

Ruiz-Santaquiteria, J., Velasco-Mata, A., Vállez, N., Bueno, G., Alvarez-Garcia, J., and Deniz, O. (2020). Handgun detection using combined human pose and weapon appearance.

Slobogin, C. (2003). Public privacy: Camera surveillance of public places andthe right to anonymity.

University, G. (2019). What is Public Safety and Where Do You Fit in? `https://www.goodwin.edu/enews/what-is-public-safety-and-where-do-you-fit-in/`. [Online; Last accessed 15 Jun 2021].

Wilson, B., Hoffman, J., and Morgenstern, J. (2019). Predictive inequity in object detection. *CoRR*, abs/1902.11097.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.