https://doi.org/10.17163/abyaups.49.381

Enhancing Autonomous Vehicle Safety: using Convolutional Neural Networks for Police Detection

Denys A. Dután
Universidad Politécnica Salesiana, Ecuador
Orcid: https://orcid.org/0000-0002-0716-3799

Juan D. Valladolid
Universidad Politécnica Salesiana, Ecuador
Orcid: https://orcid.org/0000-0002-3506-2522

Juan P. Ortiz
Universidad Politécnica Salesiana, Ecuador
Orcid: https://orcid.org/0000-0002-9185-6044

Fabricio Espinoza
Universidad Politécnica Salesiana, Ecuador
Orcid: https://orcid.org/0000-0003-4559-4474

Abstract

According to the World Health Organization, Ecuador experiences a significant number of road accidents, resulting in numerous deaths, both among vehicle occupants and pedestrians. To address this issue and reduce fatalities, the country is actively working on implementing Autonomous Vehicles (AVs) on its streets. Unfortunately, this technology is far from perfect such is the case of Waymo in which it's vehicle's failed to respond to a law enforcement officer's command to yield, highlighting the need for improvement in AV technology. To tackle this challenge, we have applied computer vision and artificial intelligence model to accurately identify traffic officers. The methodology consists of using web crawling to collect an image dataset of traffic officers in Cuenca, Ecuador. Then procced to prepare the data set to apply to the models, in this case we used three variants of the YOLO (You Only Look Once) model, specifically YOLOv3s, YOLOv5s, and YOLOv8s and evaluated their behavior. Through experimentation, the YOLOv8s model demonstrated excellent detection capabilities, achieving its an F1 score of 0.78 at a confidence threshold of 0.907. The objective of this model is to enhance AVs' ability to accurately recognize

traffic officers, thus improving road safety. As a future enhancement for this project, the researchers plan to create a larger dataset using different images of law enforcement authorities involved in vehicular traffic management in Ecuador. This expansion aims to further improve the model's accuracy and performance.

Introduction

The World Health Organization (WHO) has published a comprehensive global status report on road safety. According to the report (World Health Organization, 2018), approximately 34 % of road traffic deaths in the region involve drivers or passengers of four-wheeled vehicles. A specific analysis focused on Ecuador, conducted by the WHO, sheds further light on the situation. The study indicates that Ecuador experiences approximately 1.1 thousand fatalities among car users and 4.2 thousand fatalities among pedestrians (World Health Organization, 2023). For this reason, Ecuador is looking for the implementation of AV's since it is gaining popularity and acceptance within society (Moody, 2020). Unfortunately, this technology is far from being perfect and still needs perfection as is the case of WAYMO (Kay, n.d.) in which a vehicle fails to detect and yield to a traffic officer thus not knowing what decision to take.

AVs' are a groundbreaking advancement in transportation technology that operate and navigate without direct human input. These vehicles integrate an array of sophisticated technologies such as multiple sensors to perceive their environment, and make informed decisions, and autonomously control their movements. The transformative potential of AVs extends beyond individual convenience, as they hold the promise of reshaping the transportation system, urban planning, and society at large (Milakis, 2019). However, it is essential to recognize that the successful integration of AVs into society requires the simultaneous development of smart cities that can support their efficient functioning and seamless integration within the urban landscape.

Deep learning, as a prominent subfield of machine learning, has garnered substantial interest and demonstrated remarkable efficacy across diverse domains. This sophisticated approach revolves around training artificial neural networks to acquire knowledge and make intelligent decisions by autonomously uncovering and extracting patterns. Notably, a defining characteristic of deep learning lies in its capacity to automatically derive meaningful representations of features from raw data (Idrovo-Berrezueta et al, 2022).

Computer vision focuses on enabling machines to comprehend and interpret visual information from digital images or videos, emulating the capabilities of human vision. One of its primary objectives is to develop algorithms that facilitate the classification of objects, object detection, image segmentation, and pose estimation. Key techniques employed in this field include feature extraction, pattern recognition, machine learning, and deep neural networks, which play pivotal roles in enhancing the understanding and analysis of visual content (Szeliski, 2022). These advancements in computer vision have wide-ranging applications and hold great potential for further advancements, paving the way for a new era of visual perception and intelligent systems.

This study focuses on the application of YOLO for traffic officer detection, an essential task for the development of intelligent transportation systems. The task to distinguish pedestrians from traffic officer would allow the vehicle to take different decision and avoid incidents such as the WAYMO case. To approach this problematic this research is divided into several sections, beginning with a review of related works in the field of object detection and recognition. The methodology section outlines the approach employed to train and fine-tune the YOLO model for accurate detection of traffic officers. The experiment and preliminary results section presents the dataset used, evaluation metrics, and performance analysis of the developed model. Finally, the conclusion summarizes the key findings, discusses the limitations of the approach, and suggests potential areas for future research, highlighting the significance of traffic officer detection in enhancing road safety and intelligent transportation systems.

Related Work

In (Rafique et al., 2023) the authors discuss the topic of traffic congestion. From this research, they determined that around 40 % of the traffic congestion is caused by cars that are cursing looking around for parking spaces. As a solution to this problem, they propose the implementation of Deep learning and they do so by using YOLO detection model. This model helps detect empty parking spaces and helps reduce time in search for parking lots, as a result of this research they compared different models such as YOLOv3, AlexNet, SVM, and YOLOv5, from which the model YOLOv5 gave them the best result with a precision of 99.5 %.

Another research that used YOLO was (Fazri & Candradewi, 2023) in which the authors proposed the use of this deep learning technique to decrease the number of road traffic accidents. To do so they trained their model to detect three different traffic violations: running through a red light, helmet violation, and wrong way. They used a video game simulation from GTA V to train their model and used a total of 25197 images and 2825 for validation. As a result of this research, they created a model with different F1-scores for each traffic violation. They obtained an 0.92 F1-score for running a red light, an 0.88 F1-score for helmet violation, and finally, they obtained 1.00 for the detection of a vehicle that is going in the wrong way.

This research (Chen et al., 2023) proposes the use of YOLOv5 for the detection of Welding Helmet Use (WHU). This research focuses on detecting that welders have appropriate equipment when welding to prevent welders from suffering occupational diseases such as respiratory problems. To approach this problem, first they detected human faces to detect when a welding helmet is not used appropriately. They also detected helmets when used correctly, ignoring any helmet that may not be in use. To experiment they used a variety of YOLO versions: YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv3, YOLOv3-tiny, WHU-YOLO. In the end, they concluded a better performance of the WHU-YOLO when applying a face-assisted training method to reduce the false positives.

The research (Allebosch, Van Hamme, Veelaert, & Philips, 2023) presents a method for detecting crossing pedestrians and other objects moving perpendicular to the direction of a vehicle. The approach involves combining video snapshots from multiple cameras arranged in a linear configuration and capturing different time instances.The proposed method achieves an F1 detection score of 83.66 % and a mean average precision (MAP) of 84.79 % on an overlap test. When combined with the Yolo V4 object detector in a cooperative fusion approach, the method significantly improves the maximal F1 scores of the detector on the same dataset from 87.86 % to 92.68 % and the MAP from 90.85 % to 94.30 %. Additionally, when combined with the lower power Yolo-Tiny V4 detector using the same approach, it leads to F1 and MAP increases from 68.57 % to 81.16 % and 72.32 % to 85.25 %, respectively.

The research (Dewi et al., 2023) focuses on convolutional neural network (CNN)-based object detection algorithms, specifically Yolo V2, Yolo V3, Yolo V4, and Yolo V4-tiny. To support the experiments, the researchers have developed the Taiwan Road Marking Sign Dataset (TRMSD) and made it publicly available for other researchers to utilize. Notably, the “No Flip” setting proves beneficial for the results obtained with Yolo V4 and Yolo V4-tiny. The best-performing model in the experiments is Yolo V4 (No Flip), achieving a test accuracy of 95.43 % and an Intersection over Union (IoU) of 66.12 %. Specifically, on the TRMSD dataset, Yolo V4 (No Flip) outperforms state-of-the-art approaches, achieving a training accuracy of 81.22 % and a testing accuracy of 95.34 %.

Here is a related work that talks about new applications for AVs and how to improve their automated driving.

Wiederer et al. (2020) the authors discuss the concept of human gestures and how traffic officers indicate certain traffic rules to drivers given certain situations. To approach this problem, they propose the use of computer vision and the use of LiDAR, and radar for the recognition of hand detection, hand tracking, and hand gesture recognition. They tested a variety of convolutional neural networks among the best were: Bi-GRU with an accuracy of 82.70 in cross-subject scenarios, Bi-LSTM with an accuracy of 84.27 in Cross-view scenarios, and finally LSTM with an accuracy of 75.62 in Real-World scenarios.

Methodology

This section presents the tools that were used including the parameters that were configured for the experiments.

Web Crawling

Web crawling, also known as web scraping, is a fundamental tool for systematically extracting data from websites on the internet. It operates by initiating visits to a predefined list of seed URLs or a specific webpage, followed by the extraction of relevant data such as text, images, and links. The web crawling process is designed to explore interconnected pages, creating a network that allows for comprehensive data retrieval. The crawling can be instructed to continue until a certain depth is reached or a specific number of pages has been extracted, providing flexibility and control over the scope of data collection. Web crawling plays a crucial role in various applications, including data mining, content aggregation, market research, and information retrieval. By automating the data extraction process, web crawling enables researchers to gather large volumes of data efficiently and analyze it for valuable insights.

Scrapy

This tool, Scrapy, serves as a powerful web crawling solution that plays a fundamental role in this research by significantly reducing the time required for dataset collection. Specifically, it was crucial to create a dataset of images featuring traffic officers in the region of Cuenca in the country of Ecuador. Scrapy proves to be user-friendly and facilitates seamless navigation through websites, including popular social media platforms such as Facebook. By harnessing the capabilities of Scrapy, researchers and developers can construct robust and efficient image extraction systems tailored for social media environments. Nonetheless, it is essential to emphasize the importance of adhering to the terms of service, privacy policies, and any limitations imposed by social media platforms to ensure the ethical and legal utilization of the extracted data. Table 1 gives a summary of how the images were divided for its training and validation process.

Table 1

Dataset of images collected using scrappy and divided into subsets

Category	Train	Valid	Test
Traffic Officers (EMOV)	39	5	5

You Only Look Once (YOLO)

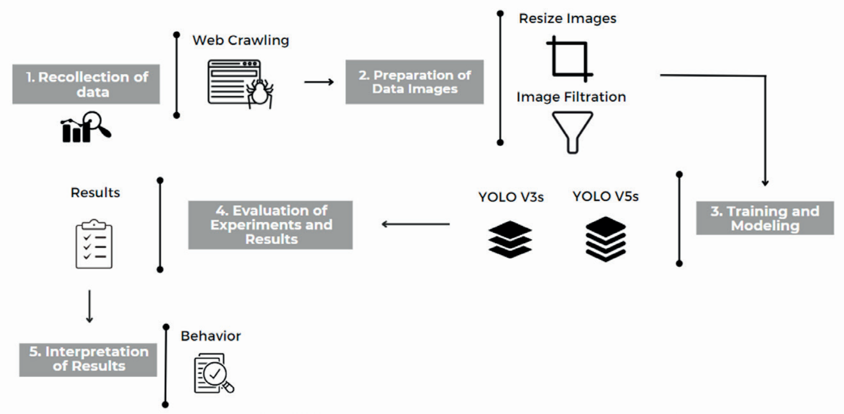
The YOLO (You Only Look Once) algorithm is a widely recognized object detection technique that has garnered significant attention due to its real-time and efficient capabilities in identifying and categorizing objects within images or videos. Unlike other object detection algorithms, YOLO processes the entire image in a single pass. It achieves this by dividing the input into a grid and predicting bounding boxes and class probabilities for each grid cell. The neural network architecture of YOLO consists of a single convolutional layer. Furthermore, YOLO offers different variations and sizes, including Tiny, S, M, L, and X, where the variation lies in the number of neurons within the convolutional layer. In this research, multiple versions of YOLO were implemented, all with the same size of S and an RGB image input of 840x840.

Experiment and Preliminary results

In our experimental and results phase, we will be referring to a designated diagram Figure. 1, which serves as a comprehensive framework for our research.

Figure 1

Training Graph



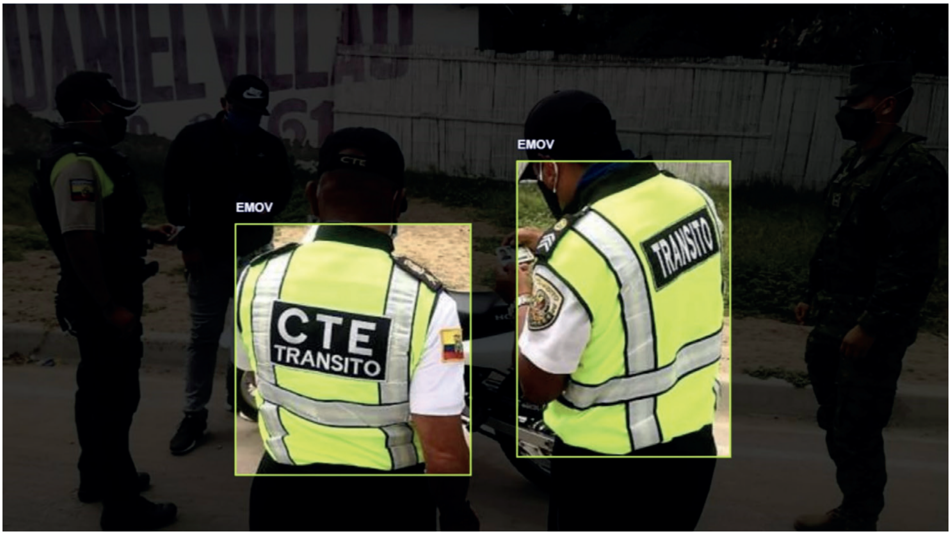
Recollection of Data

The initial step outlined in the diagram involves the collection of data through web crawling, specifically from a public platform like Facebook, due to the scarcity of existing image datasets pertaining to traffic officers. To address this limitation, we conducted a targeted search for images of traffic officers in a specific region of Ecuador. In order to expedite the process and minimize manual effort, we employed a keyword-based approach, focusing on images with descriptions containing the term "EMOV." This enabled us to gather relevant images more efficiently. After gathering the images, we proceeded to filter out irrelevant ones and retained only those depicting traffic officers. This filtering process resulted in a dataset comprising 49 images. For the purpose of training a YOLO model to accurately distinguish traffic officers from pedestrians, we assigned a single class, namely "EMOV," to the dataset.

Preparation of Data Images

Moving on to the second step, we utilized a third-party application called "roboflow" to mark the regions of interest within each image using rectangular bounding boxes and assign corresponding class labels. Since our research primarily focused on detecting traffic officer uniforms, we deemed it unnecessary for the YOLO model to learn facial recognition as seen in Figure. 2. Additionally, within the same application, we resized the images to dimensions of 840 x 840 and divided the dataset accordingly. The emphasis on detecting traffic officer uniforms in this research stems from the objective of enabling AV's navigating the streets to discern these uniforms and differentiate law enforcement authorities from pedestrians. By doing so, the AV's can make informed decisions and take appropriate actions, rather than simply avoiding or waiting for the obstacle to clear.

Figure 2
Use of Roboflow to mark the areas of interest in an image



Training and Modeling

In this section, we have chosen two distinct YOLO models: YOLO v3 and YOLO v5. Both models are characterized by their small size, indicating their low weight. For a more comprehensive overview of the applied parameters for each model, please refer to Table 2.

Table 2
Dataset of images collected using scrappy and divided into subsets

Model	Input Size	Batch Size	Epochs	Learning Rate	Momentum	Weight Decay
YOLO v3s	480 x 480	64	200	0.001	0.9	0.0005
YOLO v5s	480 x 480	16	200	0.01	0.937	0.005

Evaluation of Experiments and Results

Through training our YOLO models using the parameters outlined in Table 2, we generated a graph illustrating the F1 confidence curve. This graph demonstrates the interaction between the F1 score and the confidence coefficient as they increase. Comparing Figure 3 and Figure 4, it is evident that the YOLO v5s model achieves a higher F1 score compared to the YOLO v3s model, which yields a lower F1 score. It is important to note that a higher F1 score corresponds to a reduced number of false positives and false negatives, thus enhancing detection accuracy. The confidence coefficient plays a role in filtering predictions, allowing only objects with a certain confidence level to be visualized.

Figure 3
F1 Confidence curve from YOLO v3s

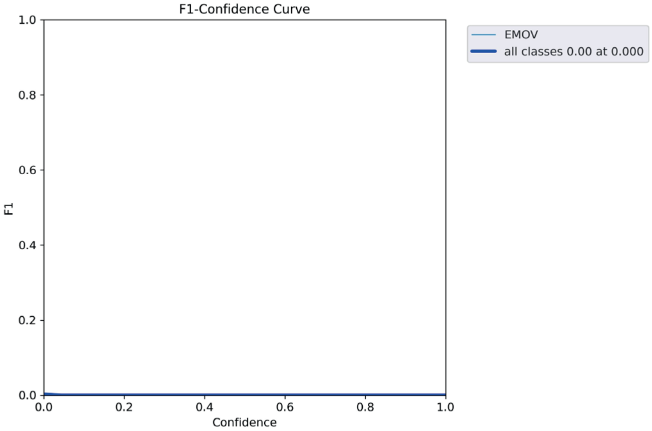
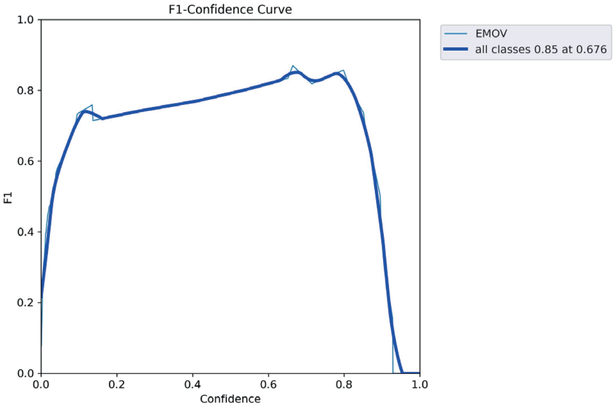


Figure 4

F1 Confidence curve from YOLO v5s



Interpretation of Results

After conducting a thorough evaluation and analysis of our models, it becomes evident that the YOLO v5s model exhibits superior accuracy. It achieves an impressive peak F1 score of 0.85 at a confidence coefficient of 0.676. An example of the behavior of our model can be seen in Figure 5, here we can visualize how the model has a variety of confidence coefficient and determine how accurately it is the detecting traffic officers. To further contextualize this finding, we have prepared a comparative analysis with the model discussed in the related works, which is presented in Table 3. It is important to consider that although our achieved F1 score may appear relatively lower compared to other models, it should be noted that this research involved the creation of a custom dataset comprising a smaller number of images. In contrast, other models examined in the related works benefited from larger datasets, which likely contributed to their higher precision and F1 scores.

Table 3

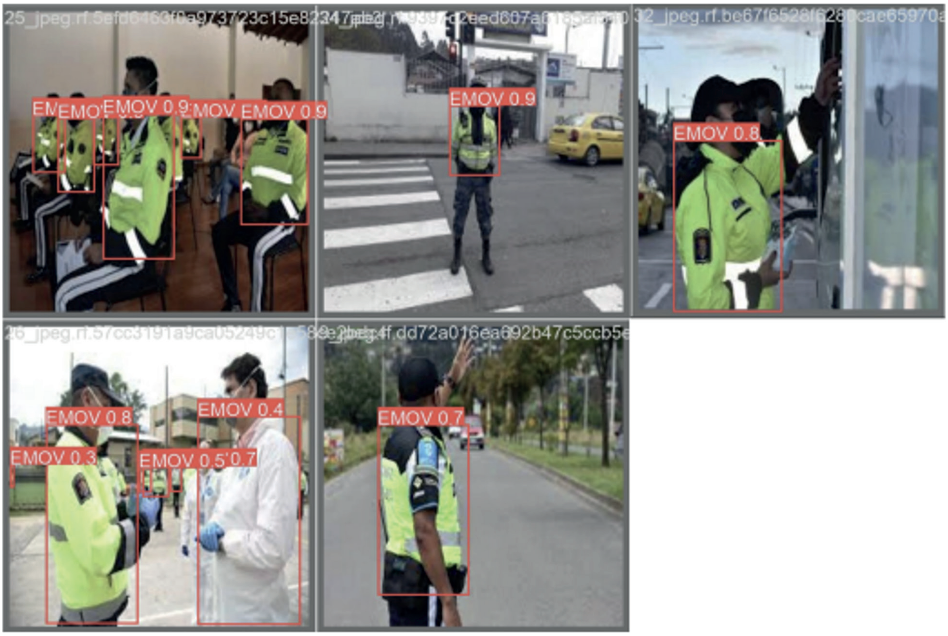
Comparative table of F1-scores of various models, Comparing the results of this research with three notable articles [1], [3], and [4], listed from top to bottom order

Model	Dataset Size	F1-Score
YOLO v5s	49	0.85
YOLO v4	NA	0.8525
YOLO v4	6009	0.89
YOLO	25197	0.92

Conclusion

Based on the insights gathered from Table 3, we can draw meaningful interpretations regarding the reasonable F1 score achieved considering the size of our dataset. An exemplary demonstration of our model's detection capabilities can be observed in Fig. 5. This image highlights the model's precision levels and the varying confidence coefficients across different scenarios. Despite the limited size of our dataset, our model exhibits immense potential for future implementation in AVs. It underscores the significance of computer vision in AVs, particularly its ability to effectively detect and differentiate pedestrians. This research aims to contribute to the advancement of AV technology, and we propose future investigations that involve expanding the dataset to include encounters with various law enforcement entities. By considering not only traffic officers but also police officers and military personnel, we can further enhance the comprehensiveness and applicability of our research.

Figure 5
Result detection of YOLO V5s



Reference

Chen, W., Li, C., & Guo, H. (2023). A lightweight face-assisted object detection model for welding helmet use. *Expert Systems with Applications,* 221.

Dewi, C., Chen, R., Zhuang, Y., Jiang, X., & Yu, H. (2023). Recognizing road surface traffic signs based on yolo models considering image flips. *Big Data and Cognitive Computing,* 7 (1). https://n9.cl/ljgf2/

Fazri, I., & Candradewi, I. (2023). "Integrated traffic violation type detection and recognition system using video processing based convolutional neural network". *ICIC Express Letters, 17*(5), 595-604.

Idrovo-Berrezueta, P., Dután-Sánchez, D., Hurtado-Ortiz, R., & Robles-Bykbaev, V. (2022, November). "Data Analysis Architecture using Techniques of Machine Learning for the Prediction of the Quality of Blood Fonations against the Hepatitis C Virus". In *2022 IEEE International Autumn Meeting on Power, Electronics and Computing* (ROPEC) 6, pp. 1-7. https://n9.cl/q82hs/

Kay, G. (n.d.). *Video shows passengers in driverless car reacting to police officer gesturing to pull over: "We're sorry, this car won't let us move!".* https://n9.cl/9xs2d

Milakis, D. (2019, January). "Long-term implications of automated vehicles: an introduction. *Transport Reviews", 39*(1), 1–8. https://n9.cl/vsfd4/

Moody, J., Bailey, N., & Zhao, J. (2020, January). "Public perceptions of autonomous vehicle safety: An international comparison. *Safety Science,* 121, 634–650. https://n9.cl/nuc34/

Rafique, S., Gul, S., Jan, K., & Khan, G. M. (2023). "Optimized real-time parking management framework using deep learning. Expert Systems with Applications", 220.

Szeliski, R. (2022). "Computer Vision: Algorithms and Applications. Springer Nature". https://n9.cl/vlenup/

Wiederer, J., Bouazizi, A., Kressel, U., & Belagiannis, V. (2020). "Traffic control gesture recognition for autonomous vehicles". https://n9.cl/rfxvy/

World Health Organization. (2018). "Global status report on road safety 2018". https://n9.cl/t634d0/

World Health Organization. (2023). "Death on the roads". https://extranet.who.int/roadsafety/death-on-the-roads/countryorarea/ECU