INSTITUTE FOR HOMELAND SECURITY

Sam Houston State University

Deep Learning

Based Traffic Accident Detection

Institute for Homeland Security

Sam Houston State University

Fan Liang

Deep Learning Based Traffic Accident Detection

Fan Liang
Department of Computer Science
Sam Houston State University
Huntsville, TX
Fxl027@shsu.edu

Abstract—In order to improve road safety, this project creates a deep learning-based system for identifying traffic incidents in pre-recorded video recordings. The project employs the YOLOv5 model, which utilizes advanced convolutional neural networks for efficient and accurate object detection. Trained on a focused dataset of video-derived images, the system demonstrated recognizing traffic incidents. From the Highway Incidents Detection (HWID12) dataset, 50 video segments were selected for this study and analyzed with different variety of the Yolov5 parameters. The YOLOv5x variant reached the highest precision of 0.6 when trained with a batch size of 4 and image resolution of 640 pixels, indicative of the model's adeptness at discerning relevant features in complex visual data. While the overall highest recall was 0.45 by the YOLOv5s model, the YOLOv5l variant demonstrated the best balance between precision and recall with the highest F1 score of 0.44. These results, although preliminary, demonstrate the potential of convolutional neural networks in improving road safety and set a precedent for the application of such technologies in broader public safety measures. Despite the limited dataset, the results show the feasibility of using deep learning techniques in traffic safety applications. Providing innovative approaches to traffic management and road safety, this study sets the stage for further research and the potential broader application of such technologies in public safety measures.

Index Terms—Deep learning, Smart Transportation

1 Introduction

The increasing number of motor vehicle accidents worldwide presents a significant challenge, primarily due to human errors, delayed emergency responses, and secondary accidents. Studies indicate that human error accounts for over 90% of all traffic accidents [1]. These accidents not only cause devastating loss of life but also contribute to substantial socio-economic costs. Recent statistics reveal a slight decrease in annual fatalities to 1.19 million worldwide, yet road traffic injuries continue to be the leading cause of death among young individuals aged 5-29 years. The global cost of road traffic crashes is estimated to be $518 billion per year, which represents between 1-2% of the Gross National Product (GNP) of each country. For low- and middle-income countries, the cost can be as high as 5% of GNP [2]. As the global vehicle population continues to rise, the potential for accidents escalates correspondingly, highlighting the urgent need for improved road safety measures.

Current methods for traffic accident detection primarily rely on manual observation and

report-based systems, which are time-consuming and often inaccurate. Moreover, existing automated systems struggle with high false positive rates and limited ability to analyze complex traffic scenarios dynamically. The reliance on traditional surveillance and reporting methods fails to leverage the full capabilities of modern technology, thus delaying response times and inhibiting efficient traffic management. Additionally, many of the automated systems suffer from biases inherent in the datasets used for training ma- chine learning models. These datasets often do not represent the diversity of real-world scenarios, leading to models that perform poorly in unrepresented or underrepresented environments. For example, systems trained primarily in well-structured urban settings may not perform effectively in rural or less structured environments. Furthermore, the variability in weather conditions, lighting, and road types poses additional challenges that current technologies struggle to handle effectively. This limitation restricts the applicability and reliability of these systems across different geographic and environmental conditions, which is critical for achieving global road safety improvements.

This project proposes a novel approach to traffic accident detection using advanced deep learning model YOLOv5 and its several pre-trained models, applied to pre-recorded video data from the Highway Incidents Detection (HWID12) dataset [3]. By automating the detection and analysis of traffic accidents, the project aims to significantly reduce the time between an accident occurrence and its detection. The integration of these models allows for real-time, accurate identification and analysis of traffic incidents, potentially transforming how traffic management systems respond to accidents.

To summarize, we make the following contributions in our project:

- The application of YOLOv5 and its several pre-trained models for accident detection represents a significant advancement over traditional detection methods, improving accuracy and reducing false positives by effectively analyzing complex motion patterns and temporal dynamics in video data.
- By automating the analysis of traffic incidents, this project contributes to the development of more responsive traffic management systems, which can lead to quicker emergency responses and better-informed traffic safety measures.
- The methodologies developed in this project demonstrate potential scalability and adaptability to different types of traffic environments and conditions, paving the way for broader applications in global traffic safety improvement efforts.

The remainder of this paper is structured as follows: Section II related works to traffic accident detection and deep learning applications in traffic management. Section III details the methodology, including data collection, data pre-processing, model training, and system implementation. Section IV presents the performance evaluation. Section V discusses the implications of these findings and suggests directions for future research. Finally, Section VI concludes the paper by summarizing the key outcomes and contributions of this project.

2 Related Works

The increase in traffic incidents, coupled with their significant impact on public safety, has prompted an urgent call for innovative solutions. Recent advancements in the fields of deep learning and computer vision have emerged as promising avenues to meet this challenge, as demonstrated by several pivotal studies.

Kyu Beom Lee's research [4] introduces a notable Object Detection and Tracking System (ODTS) that integrates deep learning with object tracking to automate the detection of accidents within tunnel environments. This system showcases a robust capability in identifying incidents such as wrong-way driving, sudden stops, and fires using CCTV footage. Despite its efficacy, the study acknowledges limitations, particularly in the area of false positives for fire detection, indicating room for improvement in the accuracy of environmental anomaly detection.

Another study [5] explores the application of gated recurrent unit (GRU) and convolutional neural network (CNN) models, employing an ensemble technique that leverages both video and audio data from dashboard cameras. This innovative approach aims to enhance the predictive accuracy of models by using a combination of weighted averages and multiple classifiers on diverse data inputs. While promising, the research focuses on singular data streams, suggesting a potential gap in exploiting the full spectrum of multi-modal data for comprehensive incident analysis.

Furthering the discourse, Arifeen et al. [6] propose a sophisticated framework for traffic accident detection and classification, utilizing deep neural networks. This framework employs convolutional neural networks (CNNs) like GoogLeNet, AlexNet, and VGGNet for initial detection, followed by anomaly classification using One-Class Support Vector Machines (OCSVM) and multi-class SVM models. Conducted on the UCF-Crime Road accident video sequences, their study demonstrates notable success in accurately detecting and classifying traffic accidents, underscoring the efficacy of deep learning and SVMs in this critical domain.

The study by K. A. D. Devindu Dharmadasa and colleagues [7] proposes an advanced road accident detection system using YOLOv5 and StrongSort for analyzing highway CCTV footage. This model, aimed at enhancing road safety, focuses on detecting vehicle accidents through vehicle speed, acceleration, and trajectory anomalies. It shows promising results in identifying accidents under various conditions, such as day- light and night. Despite its initial success, the system's reliance on specific parameters for accident prediction hints at potential areas for refinement and future development.

3 The Methodology

The methodology section of this project delineates the comprehensive approaches undertaken for effective traffic accident detection using deep learning technologies. It encompasses detailed phases of data collection, frame extraction and annotation, and the deployment of the YOLOv5 model, outlining each step involved in preparing, processing,

and analyzing the video data from the Highway Incidents Detection (HWID12) dataset.

3.1 Dataset Collection

In our project, we used The HWID12 dataset, introduced by Kezebou, Landry, Victor Oludare, Karen Panetta, James Intriligator, and Sos Agaian, which was specifically designed for Intelligent Transportation Systems (ITS) applications. Comprising over 2,780 video segments and more than 500,000 temporal frames, this extensive collection spans 11 distinct incident categories along with additional categories for negative samples representing normal traffic conditions. This dataset's comprehensive nature addresses the critical need for highly annotated data suitable for training models to detect and classify highway incidents effectively.

For the scope of this final project, due to constraints in time and complexity, our study focused on a curated selection of 50 video segments from the HWID12 dataset. This subset was chosen to allow for a detailed application and evaluation of advanced action recognition models within a feasible framework. By concentrating on these segments, we were able to thoroughly explore and test the capabilities of our models to differentiate between normal traffic conditions and various types of highway incidents, demonstrating the potential of our methodologies in real-world settings.

3.2 Frame extraction and annotation

In this project, a rigorous, hands-on technical approach was adopted for frame extraction to ensure the utmost precision in our traffic accident detection system. Frames were meticulously extracted from the raw dataset by implementing a manual selection process rather than relying on automated tools. Specifically, one frame was extracted for every ten frames to construct a dataset that balances granularity with manageability.

This manual extraction was instrumental in capturing the subtle nuances of each incident, which automated software might overlook. For the annotation of these frames, we utilized the VGG Image Annotator (VIA) [8], a versatile tool that allowed for precise demarcation and labeling of traffic incidents. Through VIA, each frame was carefully examined, and bounding boxes were manually drawn around vehicles and other relevant entities involved in traffic incidents. Each incident was then annotated with a 'traffic accident' label, which involved a keen eye for detail to ensure that only relevant objects were marked, reducing noise and potential inaccuracies in the dataset.

Figure 1: VGG Image Annotator

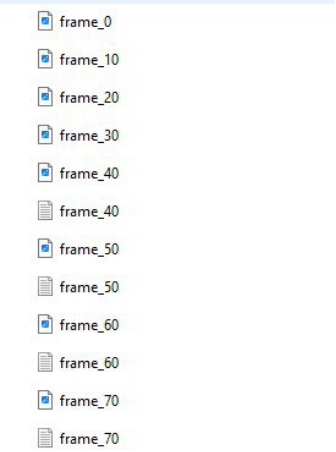

Figure 2: File Structure

The outcome of this annotation phase was a JSON file containing all the metadata for the annotated frames. To make this data compatible with the YOLOv5 model, a Python script was employed to convert the JSON output into a .txt format that YOLOv5 could process. This script not only transformed the data format but also preserved the integrity of the annotations, ensuring that the object detection model could be trained effectively with high-quality, well-labeled data.

As for the figures 1 and 2, they serve as exemplars of the annotation process. Figure 1 is be used to illustrate the environment within the VIA tool, showcasing the interface and the manual annotation in progress. Figure 2 is employed to demonstrate the detailed file structure of the extracted frames, highlighting the methodical frame selection and organization process. Both images will visually support the explanation of the data preparation stage, offering readers a concrete glimpse into the workflow and techniques used in the project.

3.3 YOLOv5 Architecture and Pre-trained Models

The YOLOv5 architecture is tailored for object detection, starting with an input layer that processes image and directs them to a backbone composed of CBS (Convolution + BatchNorm + SiLU) and C3 modules derived from CSPNet. These elements are critical for initial feature extraction and enhancing the network's ability to capture complex features while avoiding gradient information redundancy. An SPPF module at the end of the backbone further enhances spatial feature representation, crucial for recognizing objects at varied scales. The processed features are then consolidated

by a feature fusion network (neck), generating three distinct feature maps (P3, P4, P5) with sizes of 80x80, 40x40, and 20x20, to detect small, medium, and large objects, respectively in Figure
3. These maps are analyzed in the prediction head, where bounding-box regression and confidence assessments occur for each pixel based on predefined anchors. The result is a multi-dimensional array of object data, which is then refined through threshold setting and non-maximum suppression (NMS) in the post-processing phase to deliver precise object detection results.

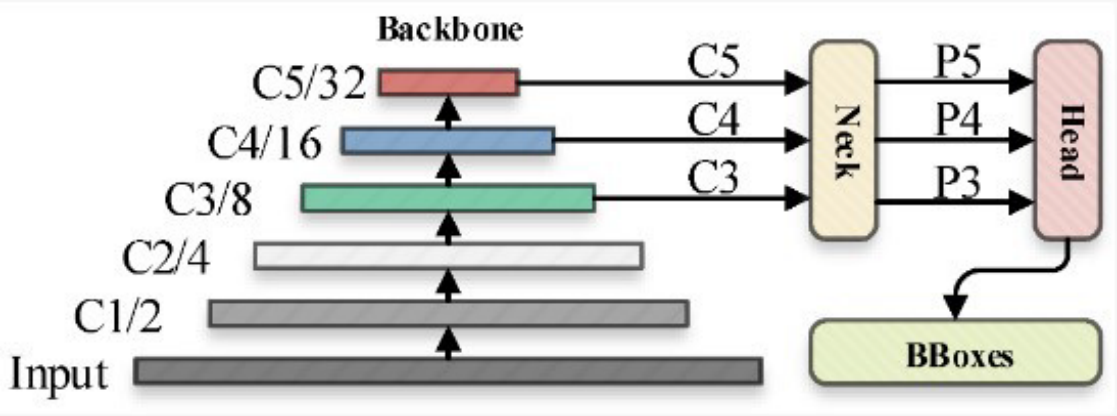

Figure 3: The default inference flowchart of YOLOv5

In our project, the YOLOv5 framework was employed with different pre-trained weights—yolov5s.pt, yolov5m.pt, yolov5l.pt, and yolov5x.pt—to identify the optimal variant for our specific application. Through iterative training and weight adjustment, we minimized the error on the training dataset, while carefully monitoring performance on the validation dataset to mitigate overfitting. The best-performing model variant, selected after rigorous evaluation, was then deployed for inference on the test dataset, yielding the final output results. This trained model, demonstrating satisfactory performance, was saved, marking a significant step towards the deployment of an efficient traffic monitoring system.

3.4 Model Training

3.4.1 Hardware Configuration

The training of our YOLOv5 model was conducted on a robust machine featuring an Ubuntu 20.04.4 LTS operating system, powered by an Intel® Core™ i9-10900X CPU @ 3.70GHz, and equipped with 62 GiB of RAM. With x86-64 architecture and support for advanced virtualization and vector extensions such as VT-x, AVX, AVX2, and AVX-512, the machine was well-suited for high-performance deep learning tasks.

3.4.2 Software and Training Procedure

For the training procedure, the dataset, consisting of 1000 frames from 40 videos, was used for training, representing 80% of the data. The remaining 200 frames from 10 videos

formed the validation set, accounting for 20% of the data. Different training epochs—100, 200, and 300—were experimented with, alongside varying batch sizes of 4, 8, 16, 24, and 32, to find the optimal balance between model learning and computational efficiency.

Hyperparameter tuning was addressed through data augmentation, implemented using PyTorch's torchvision library. The augmentation pipeline included random horizontal flips with a probability of 0.5 and standard normalization of image tensors. The augmentation aimed to improve the model's robustness to variations in real-world conditions. Various image resolutions were also tested, including 416x416, 640x640, 720x720, 1080x720, and 1280x720, to assess the impact of input size on the model's performance. These comprehensive training and tuning processes enabled the YOLOv5 model to effectively learn and accurately classify traffic accidents from video frames, setting the stage for reliable real-world application.

3.5 Data Flow and Processing

The project framework Figure 4 presents the process of training our deep learning model for traffic accident detection. Traffic accident videos are processed through VGG Im- age Annotation for frame-by-frame annotation, creating detailed labels for training the model. These frames are then divided into training, validation, and test datasets, ensuring a robust evaluation of the model's performance.

The YOLOv5 framework is then employed to train the model using different pre-trained weights (yolov5s.pt, yolov5m.pt, yolov5l.pt, and yolov5x.pt) to find the optimal

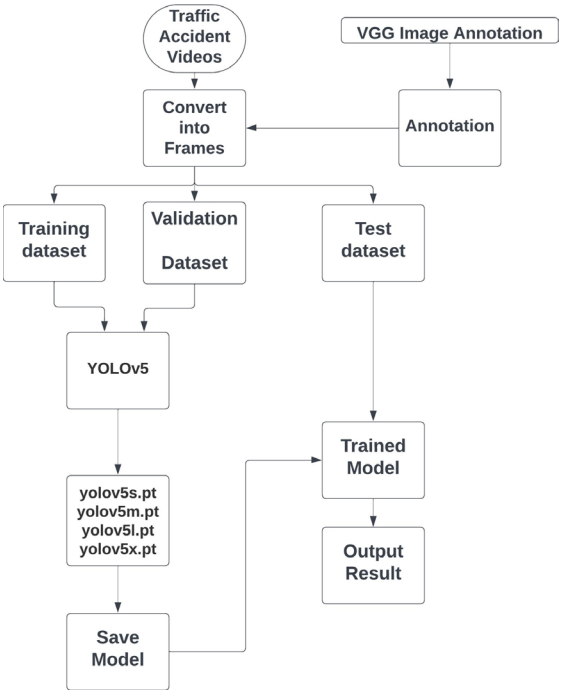

Figure 4: Project Framework

balance of speed and accuracy. The training process iteratively adjusts the weights to

minimize error on the training dataset while validation dataset performance is monitored to prevent overfitting.

After training, the best-performing model is selected and used to make predictions on the test dataset, leading to the final output result. The output is evaluated to assess the model's accuracy in detecting traffic accidents. The trained model, once satisfactory, is saved for future application or further refinement.

4 Performance Evaluation

This section assesses the performance of our YOLOv5 models using key metrics relevant to object detection, specifically the accuracy and reliability of traffic accident detection.

4.1 Evaluation Matrices

Recall is a critical evaluation metric in object detection, particularly when it is imperative to identify as many actual positive cases as possible. In the context of a traffic accident detection model, recall measures the model's ability to capture all actual accidents within the dataset. The importance of high recall in this setting cannot be overstated, as failing to detect an actual traffic accident could have severe safety implications, including missed opportunities to provide timely assistance and potentially prevent further consequences from the accident. The equation for recall is given by:

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}} \tag{1}$$

Where, True Positives (TP) are the number of traffic accidents correctly identified by the model, and False Negatives (FN) are the accidents that occurred but were not detected by the model. A high recall value indicates that the model is effective at identifying most of the actual accidents, minimizing the risk of undetected incidents.

Precision is an essential evaluation metric in object detection models, particularly in scenarios where the correctness of each positive prediction is critical. For a traffic accident detection model, precision measures the accuracy of the model in identifying true accidents, thereby indicating the likelihood that a detected accident by the model is indeed a real accident. The equation for precision is given by:

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}} \tag{2}$$

Where, True Positives (TP) are the number of traffic accidents correctly identified by the model as accidents. False Positives (FP) are instances where the model incorrectly identifies a non-accident situation as an accident.

4.2 Results

4.2.1 Small Model Parameters (YOLOv5s.pt)

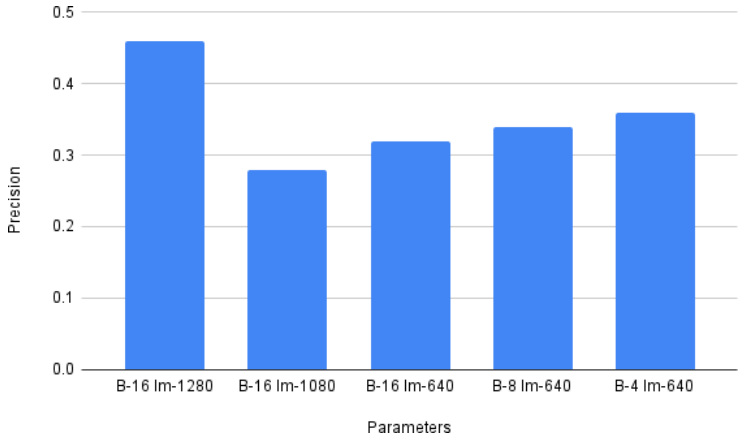

Figure 5: YOLOv5s.pt Parameters

For the smallest model, YOLOv5s, the highest precision was achieved with a batch size of 16 and an image size of 1280x720 figure 5. However, the precision differences across various configurations were marginal, indicating that the model's smaller size might limit its capability to differentiate between resolutions and batch sizes as effectively as its larger counterparts.

4.2.2 YOLOv5l and YOLOv5m Parameter Influence

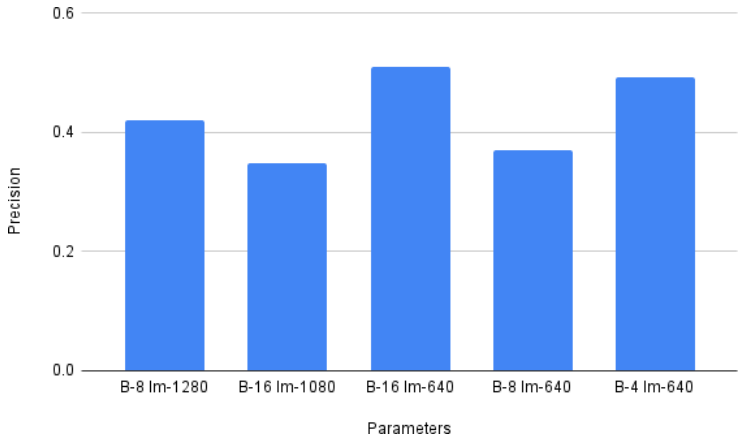

Figure 6: YOLOv5m.pt Parameters

When assessing the YOLOv5l and YOLOv5m variants, it was observed that varying the batch sizes and image resolutions had noticeable effects on precision. For YOLOv5l, larger image sizes such as 1280x720 did not necessarily translate to higher precision figure 7, which could be indicative of the model's sensitivity to overfitting with higher-dimensional input. On the other hand, YOLOv5m maintained a more consistent precision across the parameters, suggesting robustness in feature extraction at varying scales figure 6.

4.2.3 Parameter-Specific Precision Analysis (YOLOv5x.pt)

A deeper dive into the YOLOv5x variant showcased that the highest precision was achieved with a batch size of 4 and an image size of 640x640 figure 8. Larger image sizes and different batch sizes showed a decline in precision, implying that this specific combination of batch size and image resolution might be better suited for capturing the nuances of traffic accidents within the constraints of the dataset used.

4.2.4 Model Variant Comparison

The 'Highest Precision vs. Model' graph illustrates that the largest YOLOv5 model variant, YOLOv5x, achieved the highest precision, followed closely by YOLOv5l. While the smaller models, YOLOv5m and YOLOv5s, presented lower precision scores, they still performed commendably considering their reduced computational complexity figure 9. This pattern suggests that the increased capacity of larger models allows them to capture more detailed features, which is crucial for the accuracy in detecting traffic accidents.

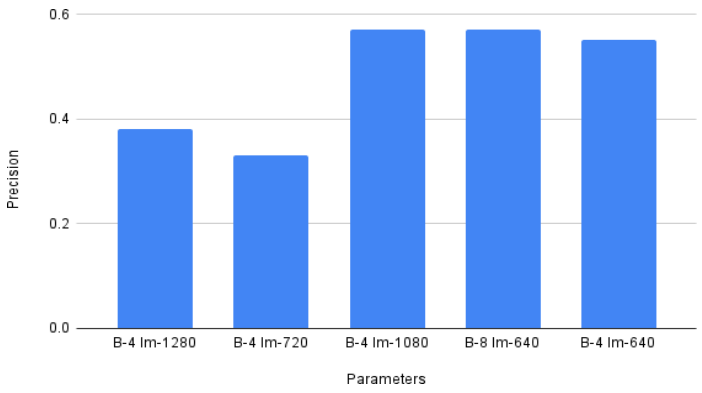

Figure 7: YOLOv5l.pt Parameters

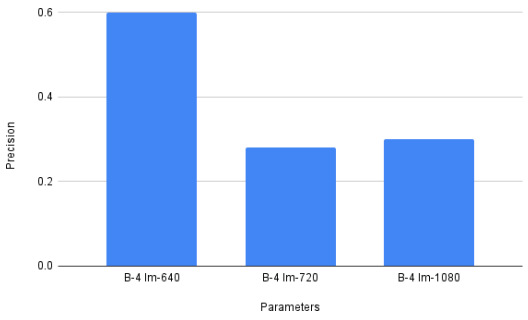

Figure 8: YOLOv5x.pt Parameters

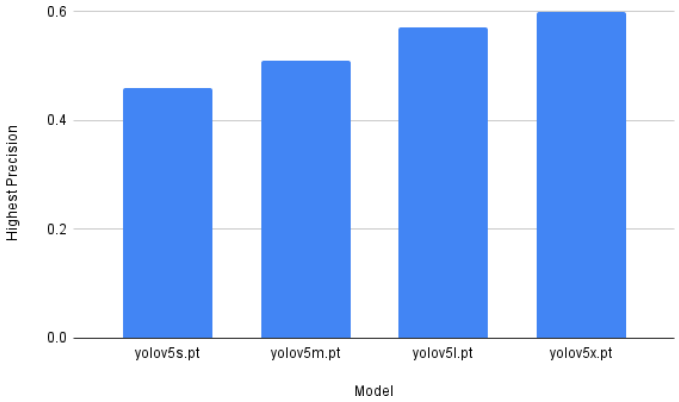

Figure 9: Model Variant Comparison

5 Discussion

5.1 Methodology Reflection

Our methodology for detecting traffic accidents using various configurations of the YOLOv5 models has demonstrated significant effectiveness. This systematic approach, under- pinned by the diverse computational capabilities of the YOLOv5 variants (s, m, l, x), facilitated a comprehensive exploration of accuracy and computational efficiency, allowing for an informed selection of the most appropriate model based on precision needs and resource availability. An annotated dataset was paramount to enhancing model accuracy, emphasizing the critical role high-quality data plays in machine learning. In addition, our experiments with varying image sizes and batch sizes have provided in- sight into how they affect model performance, resulting in an optimal balance between processing time and detection accuracy. Despite that, there is still scope for improving model robustness and generalizability, thereby mitigating overfitting issues observed in the data. This can be achieved primarily by diversifying data and increasing volume.

5.2 Limitations:

The primary limitation of this study lies in the constrained size and diversity of the dataset, which consists of only 50 videos. This limitation significantly restricts the model's ability to generalize across varied environmental conditions such as differing lighting, weather, and traffic densities—factors crucial for real-world applicability. Furthermore, the reliance on manual annotation, while ensuring data quality, introduces a risk of subjective bias and human error, potentially affecting the consistency of the training data. Additionally, focusing exclusively on traffic accidents narrows the model's application range, overlooking the detection of non-accident-related traffic behaviors that could enrich traffic management and safety strategies. To advance the applicational scope and accuracy of our models, we must address these limitations through expanded datasets and improved annotation methodologies.

6 Final Remarks

6.1 Conclusion

Through the application of different YOLOv5 models (YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x), we demonstrated a robust framework for detecting traffic accidents.

Despite the models' ability to process and analyze complex traffic incident data, the limited dataset size—only 50 videos—restricted their ability to achieve higher precision and recall. Furthermore, this limitation limited the models' ability to generalize to diverse and unpredictable scenarios, which impacted their effectiveness. Nonetheless, the use of high-quality, manually annotated data enhanced the detection capabilities of the models significantly within the constrained dataset as a result of the use of high-quality, manually annotated data.

6.2 Future Work

This research requires a significant expansion of the dataset. A larger and more diverse set of videos will enhance the models' adaptability and reliability across different traffic conditions and accident scenarios. YOLOv5 models can also be integrated with SlowFast architecture to create a hybrid model that takes advantage of both YOLOv5's capabilities in temporal analysis and location detection. We will be able to identify traffic accidents more precisely and consistently as the incident develops according to this new method, which will increase detection precision and consistency. These advances will not just improve the models' performance additionally make a major effect on the development of increasingly complex and efficient traffic safety systems.

REFERENCES

[1] World Health Organization. *Global Status Report on Road Safety 2018*. Nonserial Publication. World Health Organization, 2019.

[2] World Bank. The high toll of traffic injuries: Unacceptable and preventable. Transport Note TRN-4, World Bank, 2019.

[3] Landry Kezebou, Victor Oludare, Karen Panetta, James Intriligator, and Sos Agaian. Highway accident detection and classification from live traffic surveillance cameras: a comprehensive dataset and video action recognition benchmarking. In Sos S. Agaian, Vijayan K. Asari, Stephen P. DelMarco, and Sabah A. Jassim, editors, *Multimodal Image Exploitation and Learning 2022*, volume 12100, page 121000M. Inter- national Society for Optics and Photonics, SPIE, 2022.

[4] Kyu Beom Lee and Hyu Soung Shin. An application of a deep learning algorithm for automatic detection of unexpected accidents under bad cctv monitoring conditions in tunnels. In *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*, pages 7–11, 2019.

[5] Jaegyeong Choi, Chan Kong, Gyeongho Kim, and Sunghoon Lim. Car crash detection using ensemble deep learning and multimodal data from dashboard cameras. *Expert Systems with Applications*, 183:115400, 06 2021.

[6] Zain Ul Arifeen, Jang-Eui Hong, Bo-Seok Seo, and Jae-Won Suh. Traffic accident detection and classification in videos based on deep network features. In *2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 491–493, 2023.

[7] Surya Nasution and Fussy Dirgantara. Pedestrian detection system using yolov5 for advanced driver assistance system (adas). *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7:715–721, 06 2023.

INSTITUTE FOR HOMELAND SECURITY

Sam Houston State University

The Institute for Homeland Security at Sam Houston State University is focused on building strategic partnerships between public and private organizations through education and applied research ventures in the critical infrastructure sectors of Transportation, Energy, Chemical, Healthcare, and Public Health.

The Institute is a center for strategic thought with the goal of contributing to the security, resilience, and business continuity of these sectors from a Texas Homeland Security perspective. This is accomplished by facilitating collaboration activities, offering education programs, and conducting research to enhance the skills of practitioners specific to natural and human caused Homeland Security events.

Institute for Homeland Security
Sam Houston State University

Liang, F. (2024). Deep learning based traffic accident detection (Report No. IHS/CR-2024-1041). Sam Houston State Institute for Homeland Security.
https://doi.org/10.17605/OSF.IO/Q6BKH