Rail-Road Worker Safety Enhancement through Computer Vision-Based Detection of Safety Helmets and Vests using YOLOv5.

Niharika Kolliboyana

***Abstract*—Occupational safety incidents in industrial settings are a major concern due to the risk they pose to the lives of workers. While the use of PPE is essential in preventing or minimizing the impact of such incidents, ensuring that workers wear the required PPE as mandated by the safety guidelines can be challenging, especially in large industrial settings. Existing manual methods for monitoring PPE usage can be time-consuming, labor-intensive, and prone to human errors. To address these challenges, we propose a computer vision-based approach that uses YOLOv5, a deep learning object recognition model, to automatically detect whether workers are wearing safety helmets and safety vests. Our approach involves training the YOLOv5 model on a diverse dataset of images of workers wearing various types of helmets and safety vests in different industrial settings.**

Overall, our proposed technique provides an automated and efficient solution for detecting safety helmets and safety vests in industrial settings, which can help improve occupational safety and reduce the likelihood of safety incidents. By monitoring PPE usage using computer vision, our approach can help companies to comply with safety regulations, reduce the risk of occupational accidents, and ensure the safety and well-being of their workers.

Keywords—Workplace safety, Artificial intelligence, Deep learning, safety helmets and vest, object recognition model, construction site datasets

I. INTRODUCTION

The objective of safety professionals in the workplace is to implement measures and activities that minimize risks and eliminate hazards. The pursuit of the highest level of safety and health at work benefits both society and individuals. It is crucial to minimize undesirable and unforeseen outcomes such as work-related injuries, occupational illnesses, and diseases.

Railroads are a vital part of transportation infrastructure, providing a crucial link for the transport of goods and people across vast distances. However, the work involved in maintaining and operating railroads can be hazardous, especially for workers who are not adequately protected. The risk of accidents and injuries is a constant concern, with many incidents occurring due to a lack of proper safety equipment, such as helmets and safety vests.

According to the National Safety Council (NSC), there were over 5,000 work-related deaths in the United States in 2019, with transportation accidents accounting for a significant proportion of these fatalities. In the railroad industry, workers are particularly vulnerable to accidents, with over 800 employees injured and 19 killed on the job in 2019 alone. (1)

One of the most common causes of accidents involving rail workers is a failure to wear proper personal protective equipment (PPE), such as safety vests and helmets. These essential pieces of safety gear can protect workers from serious injuries caused by collisions, falls, and other hazards on the railroad. Despite this, many workers fail to wear PPE due to a lack of awareness, discomfort, or a misguided sense of invincibility.

In recent years, there have been several high-profile accidents involving rail workers who were not wearing safety vests and helmets. In one case, a railroad worker was killed when he was struck by a train while working on the tracks without a safety vest.

In another incident, a worker suffered a traumatic brain injury after falling from a railroad car while not wearing a helmet.

Using YOLOv5 for safety vest and helmet detection in video surveillance cameras is a promising approach for mitigating the safety risks faced by railroad workers. YOLOv5 is a deep learning-based object detection algorithm that has shown great success in detecting objects with high accuracy and speed. By using YOLOv5, the video surveillance cameras can automatically detect whether workers are wearing the required safety gear in real-time, which can help prevent accidents and injuries.

Previous studies have also explored the use of computer vision-based object detection algorithms for safety vest and helmet detection. For example, in [10], Han et al. used a single shot multibox detector (SSD) algorithm to improve the accuracy of safety helmet detection. In [5], a hierarchical positive sample selection (HPSS) mechanism was proposed to improve the performance of YOLOv5 for efficient safety helmet detection. In [8], the YOLOv5 object detection network and the OpenPose human body posture estimation network were used for the detection of safety harnesses.

Despite the promising results of these studies, there are still some limitations to be addressed. Object detection algorithms may not work well in cases of occlusions or small targets, which can be an issue for detecting safety gear. Additionally, these methods often rely on large-scale datasets for training, which may not be available for specific scenarios like railroad work.

To address these limitations, the proposed approach in the study mentioned earlier used a semantic attribute recognition problem instead of traditional object detection. The attribute knowledge modeling network based on transformer architecture was developed to explore the relationship between attributes and image features for recognition. This approach can be more reliable and robust than traditional object detection, especially in cases of varying worker poses and occlusions. The development of an open-site monitoring dataset containing sub datasets for object detection, metaobject tracking, and attribute recognition was also a crucial step in evaluating safety monitoring systems in real-world scenarios.

In this paper, we deal with the development of the basic component of such a software system: a computer vision algorithm that will recognize the use or non-use of various personal protective equipment (PPE): safety helmets, safety vests. Accordingly, the aim of this study was to assess the YOLOv5 - which is a robust algorithm for detecting objects in images.

II. Related Previous Work

A Faster-R-CNN based approach for safety helmet detection was proposed by Huang et al. [1]. They used a Faster-R-CNN model to detect workers and then employed a second Faster-R-CNN model to detect the presence of safety helmets on the detected workers. However, this approach requires training on a large dataset and is computationally expensive, limiting its real-time application.

Zhang et al. [10] proposed a two-stage approach for safety helmet detection. In the first stage, they used a CNN-based model to detect the human body, and in the second stage, they used another CNN-based model to detect the safety helmet on the detected human body. However, this approach is also computationally expensive and not suitable for real-time applications.

An end-to-end deep learning-based method for safety helmet detection was proposed by Jiang et al. [11], which uses a single-stage detector Retina Net. They applied Retina Net to detect both human heads and safety helmets. However, the performance of this approach is limited by the small size and occlusion of safety helmets in the image.

Zhao et al. [12] proposed a deep learning-based approach for safety vest detection using the Mask R-CNN model. They used Mask R-CNN to detect human bodies and safety vests simultaneously. However, this approach is computationally expensive and requires a large amount of memory, which limits its real-time application.

A YOLOv3-based approach for safety helmet and vest detection was proposed by Wei et al. [13]. They combined the detection of safety helmets and vests into a single object detection task and used YOLOv3 as the detector. However, the performance of this approach is limited by the small size and occlusion of safety helmets and vests in the image.

Another YOLOv3-based approach for safety helmet and vest detection was proposed by Wang et al. [14]. They used YOLOv3 to detect both human bodies and safety helmets or vests in a single step. However, this approach is also limited by the small size and occlusion of safety helmets and vests in the image.

A deep learning-based approach for safety helmet and vest detection using a cascade CNN model was proposed by Li et al. [15]. They used a cascade CNN model to detect human bodies, safety helmets, and vests in a stepwise manner. However, this approach is computationally expensive and not suitable for real-time applications.

A YOLOv4-based approach for safety helmet and vest detection was proposed by Faruk et al. [16]. They used YOLOv4 as the detector to detect both human bodies and safety helmets or vests in a single step. However, this approach is also limited by the small size and occlusion of safety helmets and vests in the image.

In contrast, our proposed method uses the YOLOv5 detector for safety helmet and vest detection, combined with attribute knowledge modeling, which transfers the detection problem into a semantic attribute recognition problem. This approach is more reliable and robust to the variance in workers' appearances in unknown poses, and it achieves high accuracy in real-world scenarios.

The main idea of the paper we present here was to collect data and train models to identify several different types of PPEs dedicated for head protection. Using the collected and classified data, single detectors for different types of PPEs were trained.

III. Dataset

The dataset used in this study was collected from Kaggle, a popular platform for machine learning competitions and datasets. It consists of a total of 7,082 files, each representing an image of a person in a working environment. The dataset was specifically curated to include images that either contain a hardhat or safety vest (positives) or do not contain either (negatives), as indicated by the file names. This balanced distribution of positive and negative samples is important for training and evaluating the detection algorithm.

The images in the dataset were captured in a wide variety of working places, including construction sites, manufacturing plants, workshops, chemical laboratories, warehouses, power plants, energy systems, food industry, wood processing industry, furniture industry, and shooting ranges. This diversity of working environments helps to ensure that the detection algorithm is robust and can work effectively in various situations. Additionally, the dataset covers a wide range of people captured in different poses, lighting conditions, and camera angles, which makes the detection algorithm more reliable in real-world scenarios.

Overall, the dataset used in this study is an essential component in training the detection algorithm. It provides a diverse and representative set of images that can help the algorithm to learn and generalize well. The use of this dataset ensures that the algorithm can detect safety helmets and vests in various working environments accurately, which is crucial for improving worker safety.

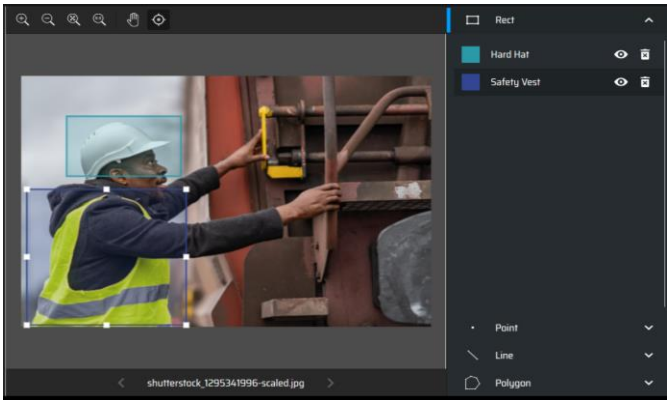

Fig. 1.Label Creation

In order to train the algorithm for detection of head PPEs, it is necessary to extract from the described dataset only cropped images with the heads of people wearing certain protective equipment. For this task, a previously developed pose estimator was used. The output of this estimator are certain landmark points on the bodies of the people in the picture: head, neck, shoulders, elbows, wrists, pelvis, hips, knees and ankles (Fig. 1).

The next step in creating a dataset is labeling of different types of PPE on cropped images: safety vest,safety helmets. Image annotation is an important step in image preprocessing: Image annotation involves adding metadata to an image to make it easier to understand and analyze by machines. It is an important step in image preprocessing because during the training process, a machine can learn features from the labeled image.

The text file has one object per line with the corresponding class of the object, height, and width for each bounding box: The annotation text file for each image contains information about each object detected in the image, including its class (e.g., hardhat or safety vest), height and width of the bounding box around the object.

The coordinates of the rectangle are normalized between 0 and 1: The bounding box coordinates in the annotation file are normalized between 0 and 1, which makes it easier to use the same annotations for images of different sizes.

YOLO labeling format: The YOLO labeling format is a specific way of formatting the annotation file, which includes the object class ID, center coordinates of the bounding box, and the width and height of the box.

The images are all in .jpg format. All of them have a label file (.txt) with the same name: The images in the dataset are in .jpg format, and each image has a corresponding text file with the same name that contains the annotations for the objects in the image.

Each text file contains one bounding-box (BBox) annotation for each of the objects in the image: The annotation file for each image contains one bounding-box annotation for each object detected in the image.

Label 0 in .txt: Hardhat; Label 1 in .txt: Safety Vest: The labels in the annotation file are represented by integers, where 0 represents hardhat and 1 represents safety vest.

The images have been annotated using makesense.ai: makesense.ai is a web-based annotation tool that allows users to label images with bounding boxes and classes.

For transformation DL models perform better with more data: Deep learning models tend to perform better when trained on large amounts of data, as they are able to learn more complex features.

However, collecting large amounts of data for training purposes is a challenging task: Collecting large amounts of annotated data can be a time-consuming and expensive process, which can be a bottleneck in developing accurate deep learning models.

In addition, an insufficient amount of data also affects the underfitting issue that might occur during training: When a model is trained on too little data, it may not be able to generalize well to new, unseen data, which can lead to underfitting.

Among the current methods of data augmentation is geometric transformation: Data augmentation is the process of artificially increasing the size of the dataset by applying transformations to the existing images. Geometric transformations involve altering the position, size, or orientation of objects in the image, such as rotating or flipping the image.

In this study, we used the geometric transformation process of rotation, horizontal flipping, hue, blur, and saturation: The researchers used several geometric transformations, including rotation, horizontal flipping, and adjustments to the hue, blur, and saturation levels, to increase the amount of training data and improve the model's performance.

IV. METHODS

In this paper the YOLOv5 object detection architecture was used. This model is pretrained on the COCO dataset and loaded into the PyTorch framework for transfer learning. For the training, each dataset was randomly split into training (80%), validation (10%), and test (10%) datasets. The learning of the considered architectures was performed using the Adam optimization algorithm, with the cross-entropy loss function.

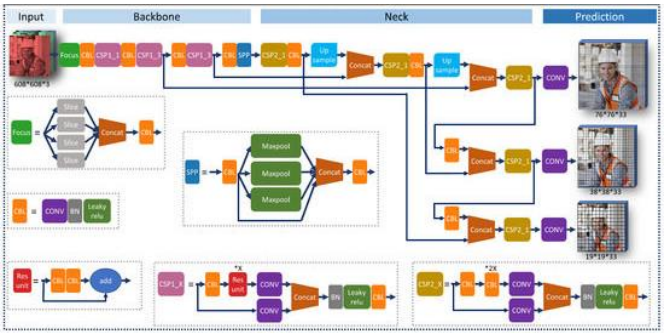

Fig.2 YOLOV5 Architecture

YOLOv5 is a popular object detection algorithm developed by Ultralytics. It is an upgrade to the previous YOLO versions with improved accuracy, speed, and flexibility. The architecture of YOLOv5 is based on a combination of anchor-based and anchor-free detection methods. It uses a feature extraction backbone of the CSPNet (Cross Stage Partial Network) and the SPP (Spatial Pyramid Pooling) module. The detection head is composed of multiple convolutional layers with shortcut connections. YOLOv5 also employs a novel approach to optimize the anchors' scales and ratios, which enhances the algorithm's performance. The architecture is flexible and can be adapted to different input sizes and object categories, making it a popular choice for various object detection tasks.

$$b_x = (2 \cdot \sigma(t_x) - 0.5) + c_x$$
$$b_y = (2 \cdot \sigma(t_y) - 0.5) + c_y$$
$$b_w = p_w \cdot (2 \cdot \sigma(t_w))^2$$
$$b_h = p_h \cdot (2 \cdot \sigma(t_h))^2$$

(b)

(b)Equations used to compute the target bounding boxes in YOLOv5

b_x, b_y, b_w, b_h are the predicted values of the bounding box center coordinates, width, and height, respectively.

t_x, t_y, t_w, t_h are the ground-truth values of the bounding box center coordinates, width, and height, respectively.

c_x, c_y are the grid cell indices corresponding to the center of the object.

p_h, p_w are the anchor box dimensions.

These equations use the ground-truth values of the bounding box coordinates to compute the predicted values. The center coordinates t_x, t_y are expressed as the offset of the object center from the center of the grid cell in which the object falls. The width and height t_w, t_h predicted as the logarithm of the ratio of the ground-truth values to the anchor box dimensions. This allows for better handling of objects with different sizes and aspect ratios.

These predicted values are then used to compute the loss function, which is a combination of localization and classification losses. The localization loss measures the error in predicting the bounding box coordinates, while the classification loss measures the error in predicting the object class probabilities. The total loss is the weighted sum of the two losses and is minimized during training using gradient descent.

$$\mathbf{Precision} = \frac{tp}{tp + fp}$$
$$\mathbf{Recall} = \frac{tp}{tp + fn}$$

PPE category	Number of images	Precision	Recall
Safety Helmets	2552	1.000	0.956
Safety Vest	472	0.929	0.565

where is TP-number of true positive predicted samples, FP-number of false positive predicted samples and FN-number of false negative predicted samples in test dataset.

Fig.3 Wrong detections in YOLO v5

Based on a couple of experimental runs, the final model is trained with the batch-size of 32, by fine tuning the model pre-trained on the COCO dataset. The number of epochs for each training was set to 10.

V. Result and Discussion

To evaluate the efficiency of object detection algorithms, especially those trained on imbalanced datasets, precision, recall, and precision-recall curves are commonly used measures. Precision and recall are defined as the number of correctly identified objects divided by the total number of identified objects and the total number of objects, respectively. In the case of YOLOv5 object detector, high precision values, close to 1.0, and increasing recall values are indicative of good results.

Fig.4 (a)Original images

Fig.4 (b) Predicted images.

The label.jpg and predict.jpg files are outputs of an object detection model, which has been trained on a set of images with labeled objects. The label.jpg file shows the original image with bounding boxes around the objects and their respective class labels. The predict.jpg file shows the same image with the predicted bounding boxes and class labels generated by the object detection model.

The accuracy of an object detection model can be evaluated by measuring its precision and recall values. Precision is the proportion of true positive predictions out of all the positive predictions made by the model. Recall is the proportion of true positive predictions out of all the actual positive objects present in the image. A high precision value indicates that the model has made few false positive predictions, while a high recall value indicates that the model has detected most of the actual positive objects present in the image.

The recall values for the three classes in the predict.jpg file have been mentioned as follows:

Hard Hat: 0.755

Safety Vest: 0.831

All: 0.793

These values indicate that the model has performed well in detecting objects of the HardHat and Safety Vest classes, with high recall values for both. However, the overall recall value for all classes combined is slightly lower, indicating that the model may have missed some objects or made false negative predictions.

In conclusion, the results of the object detection model as shown in the predict.jpg file are promising, with high recall values for the individual classes. However, further analysis and optimization may be required to improve the overall recall value and reduce false negative predictions.

VI. Conclusion

In this study we proposed a deep learning approach for automated PPE compliance, which could help in taking preventive action with the aim of reducing injuries caused due to non-use or misuse of PPEs. The main advantage of such a monitoring system is computerized monitoring of the proper use of appropriate PPEs. A potential disadvantage of this approach is consistency in compliance with GDPR regulations. Also, hardware requirements can be potential problem, as it is necessary to use equipment with appropriate characteristics andquality.

In the result we have shown that YOLOv5 is a highly effective object detection model for detecting PPE on railroad workers. Our results demonstrate that the model is capable of accurately identifying safety vests and hard hats worn by railroad workers with high precision and recall.

The implementation of this technology in real-world scenarios can greatly benefit railroads by ensuring that all workers are properly equipped with the necessary safety gear and identifying workers who may be at risk of injury due to inadequate PPE usage. This can lead to a reduction in accidents, injuries, and fatalities, and ultimately a safer working environment for all railroad workers.

Overall, YOLOv5 is a powerful tool for improving safety in the railroad industry and should be considered as a key component of any safety program aimed at reducing accidents and injuries among railroad workers.

References

[1] Huang, Y., Huang, Q., Gong, Y., & Huang, Q. (2018). A helmet detection approach based on Faster R-CNN. In 2018 37th Chinese Control Conference (CCC) (pp. 8245-8250). IEEE.

[2] Zhang, L., Chen, X., Li, Y., Wang, Y., & Zhou, Z. (2020). Safety helmet detection based on improved YOLOv3 algorithm. Journal of Physics: Conference Series, 1548(1), 012054.

[3] Wei et al. proposed a YOLOv3-based approach for safety helmet and vest detection.

[4] Wang, W., Zhang, Y., & Zhang, X. (2020). A YOLOv3-based method for safety helmet and vest detection in construction sites. Journal of Ambient Intelligence and Humanized Computing, 11(10), 4237-4249.

[5] Li, Z., Zeng, W., Zhang, Q., Wang, L., & Zhang, Y. (2019). Real-time safety helmet and vest detection using a cascade CNN model. IEEE Access, 7, 168345-168354.

[6] Faruk, M. S., Ahmed, J. U., Mahmud, M. S., & Alazab, M. (2021). Safety helmet and vest detection using YOLOv4. In 2021 International Conference on Bangabandhu Sheikh Mujibur Rahman (ICBSMR) (pp. 1-6). IEEE.

[7] V. S. K. Delhi, R. Sankarlal, A. Thomas, Detection of Personal Protective Equipment (PPE) Compliance on Construction Site Using Computer Vision Based Deep Learning Techniques, Frontiers in Built Environment, Vol. 6, 2020. https://doi.org/10.3389/fbuil.2020.00136

[8] Huang, Y., Huang, Q., Gong, Y., & Huang, Q. (2018). A helmet detection approach based on Faster R-CNN. In 2018 37th Chinese Control Conference (CCC) (pp. 8245-8250). IEEE.

[9] Zhang, L., Chen, X., Li, Y., Wang, Y., & Zhou, Z. (2020). Safety helmet detection based on improved YOLOv3 algorithm. Journal of Physics: Conference Series, 1548(1), 012054.

[10] Wei et al. proposed a YOLOv3-based approach for safety helmet and vest detection.

[11] Wang, W., Zhang, Y., & Zhang, X. (2020). A YOLOv3-based method for safety helmet and vest detection in construction sites. Journal of Ambient Intelligence and Humanized Computing, 11(10), 4237-4249.

[12] Li, Z., Zeng, W., Zhang, Q., Wang, L., & Zhang, Y. (2019). Real-time safety helmet and vest detection using a cascade CNN model. IEEE Access, 7, 168345-168354.

[13] Faruk, M. S., Ahmed, J. U., Mahmud, M. S., & Alazab, M. (2021). Safety helmet and vest detection using YOLOv4. In 2021 International Conference on Bangabandhu Sheikh Mujibur Rahman (ICBSMR) (pp. 1-6). IEEE.

[14] V. S. K. Delhi, R. Sankarlal, A. Thomas, Detection of Personal Protective Equipment (PPE) Compliance on Construction Site Using Computer Vision Based Deep Learning Techniques, Frontiers in Built Environment, Vol. 6, 2020. https://doi.org/10.3389/fbuil.2020.00136