Northeastern University

NRCan - ViaRail Computer Vision Project

Progress Report

Sanyami Shah

Dr. Ynonne Coady

December 18, 2021

1 Introduction

Every object which is used frequently requires proper maintenance and inspection to increase its life span. Via Rail operates over 500 trains per week across several Canadian provinces and 12,500 kilometres (7,800 mi) of the track. Observing them from the air is a powerful way to monitor them. The images of the railway track are captured from various distances in different locations through drones. The project aims to identify various points of interest like is Vegetation, missing ties, broken ties present on or near the railway track.

2 Related Work

Face detection has been made insufficient progress in the VR games field, due to the lack of database of VR games. In available technology of face detection, specifically for the uncovered facial occlusion, mainly relies on the eyes features to detect faces. However, it does not work to detect a face when the face has facial occlusion. Therefore, it needs to train a model with a neural network to detect faces. The first paper(5) detected the face from the VR game where they collected the face images of the VR games, annotated the position of the face and used the YOLOv5 neural network as a single target face detection.

Automating vehicle statistics gives important data that may be utilised to forecast traffic flow. In comparison to a sensor-based method, object detection-based systems that employ computer vision have demonstrated dramatic improvements in outcomes. The methodology suggested in the study(2) uses a real-time approach to accomplish this process and is now being utilised to estimate parking space density, among other things. The study offers a four-layer parking management architecture that includes HAAR-based frame extraction from a live video stream, followed by a YOLOv2 (You Only Look Once) deep neural network technique for real-time car recognition. The third layer highlights the employment of a mechanism that counts the number of cars entering a parking place by following the path drawn by the centroid, which is then followed by a number plate recognition system that can track down mishaps. The detection system constructed using this model has been extensively tested on real-time traffic in Bangalore, yielding accuracies of close to 95 percent with video data that has been manually cross-verified, making it far more effective than sensor-based models.

The accuracy of licence plate (LP) detection and identification has substantially improved because to the rapid growth of deep learning. However, due to the low computational power of embedded devices, completing this operation in real time is problematic. This research(3) proposes a real-time licence plate detection and identification network to address this issue. They created a detection unit that can detect the licence plate's bounding box and four corner points, and then use the ROIAlign approach to extract features from the same backbone in order to accomplish licence plate identification. On the large-scale licence plate data set Chinese City Parking Dataset (CCPD), the resultant architecture, dubbed RT-LPDRnet, beats all SOTA approaches while having a quicker inference time than current methods. Our code will be made available to the general public.

For deaf and dumb persons who are unable to hear or talk, sign language is their sole means of communication. There are almost 2.5 million persons in Vietnam with hearing and speech impairments, yet there are just a few sign language interpreters. Hearing challenged persons, like everyone else, require normal communication, information,

and public services such as hospitals. Because there aren't enough sign language interpreters or effective ways for regular people to communicate with the hearing impaired, a simple technology that makes sign language accessible to everyone is needed. This research(1) shows how to recognise Sign Language Gestures using a recurrent neural network (RNN) and the Mediapipe hand tracking framework. Multi-Hand Tracking and a deep learning model that can distinguish movements by Hand Landmark Features per frame with RNN training are used to build training data from input video. The collection includes motions for the most commonly used Vietnamese words. In terms of word recognition, this model delivers quite accurate results.

With the widespread use of X-ray screening devices, intelligent detection of contrabands in X-ray screening pictures has become increasingly important. Due to the random distribution of the items, which might cause the target objects and other objects to overlap, detecting contrabands in X-ray screening pictures is a difficult challenge in the field of security detection. Traditional image processing and recognition algorithms struggle to partition X-ray security pictures into separate candidate zones containing different items. In recent years, the YOLO (You Only Look Once, a Realtime Object Detection System) Model was introduced, which gives a basic framework for directly predicting bounding boxes and class probabilities from entire photos. A YOLO-based approach is utilised to detect contrabands in X-ray screening pictures in this research(4). The results of the experiments reveal that the precision and recall rate of contraband identification against a simple backdrop are both greater than 98 percent and 94 percent, respectively. Although accuracy remains around 95 percent in a complicated context, the recall rate of particular contrabands has declined to below 70 percent.

3 Methodology

3.1 Data pre-processing

Dataset pre-processing is performed to make sure that the dataset is in the suitable format for the next step in the process.

3.1.1 Image Orientation Correction

Annotation is performed in a square or rectangular block signifying the point of interest. To make the annotation process easier and efficient, the orientation of images was corrected such that tracks are in a vertical or horizontal direction.

3.1.2 Artificial Vegetation

As the rail tracks are well maintained, there were very few images where vegetation was visible. To achieve higher accuracy, we need to train the model using a well-balanced set of images with points of interest thus artificial vegetation was added on and besides the tracks using Adobe Photoshop.

3.1.3 Artificial Missing Ties

Similarly, there were no images with missing ties thus ties were removed from some of the images with Adobe Photoshop to train the model for the scenarios where the tie is missing

3.2 Annotation

After pre-processing the data, the next step was to start annotating the images for training the algorithm. 4 points of interest that were selected for the training purpose were as follows:

- Vegetation: Each image was reviewed closely to check any vegetation near the track. If it was present at a close distance to the track, it was annotated so that algorithm gets trained accordingly.
- Missing Tie: The portion of the track was annotated where there was a missing tie.
- Broken Tie: If there was a crack on the tie or seemed broken it was annotated as a broken tie.
- Others: Any un-identified object on the tracks was classified as others.

3.3 Augmentation

One of the most fascinating aspects of computer vision is the ability to enhance your effective sample size by combining current images with random alterations. Assume you have a single snapshot of a coffee mug. After that, duplicate the snapshot and rotate it 10 degrees clockwise. You haven't accomplished much, in your opinion. However, you've more than doubled the quantity of photographs you're preparing to provide your model! Your computer vision model now has a completely different viewpoint on how that coffee mug seems.

With an existing image dataset, data augmentation can increase model performance when creating computer vision models. Image augmentation expands a dataset's size and variability, enhancing model generalizability.

After the images annotation process, different augmentation was tried to compare their effect on the accuracy of the model.

- Exposure: Adjust the gamma exposure of an image to be brighter or darker.
- Saturation: Saturation augmentation is similar to hue except that it adjusts how vibrant the image is. A fully desaturated image is grayscale, partially desaturated has muted colors, and a positive saturation shifts colors more towards the primary colors.
- Mosaic: It works by taking four source images and combining them together into one.

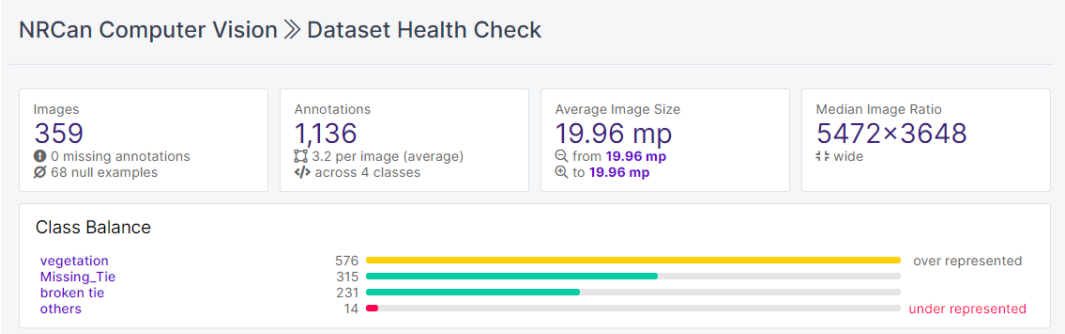

Figure 1: Dataset Health Check

3.4 Algorithm Training

Figure 1 shows the overview of the dataset. There were 359 images in the dataset, including the images containing missing or null annotations. There were 1,136 objects annotated in 359 images, out of which 576 was vegetation, 315was missing tie, 231 were broken tie, and 14 were the others.

This complete dataset was given to the algorithm where the data was split into 3 categories i.e. Training, Validation, and Testing data set. 70 percent of the images were considered under training set, 20 percent were considered under validation set and rest 10 percent were testing set.

4 Result and Analysis

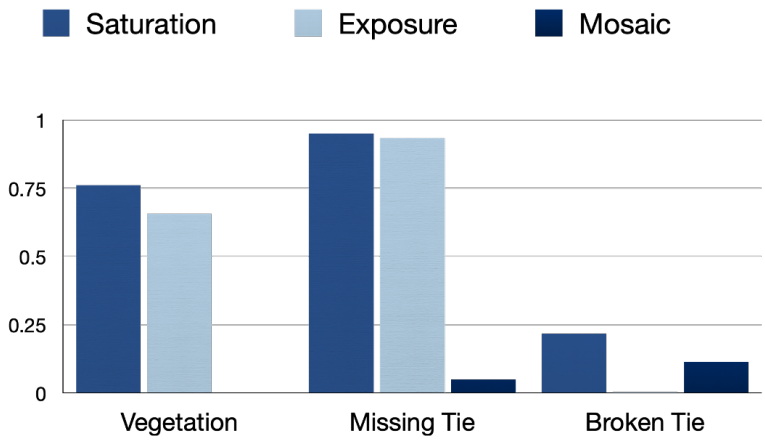

Figure 2: Model Performance Metrics

Fig 2 shows the precision result of the algorithm using different augmentation.

4.1 Saturation

4.1.1 Model Accuracy Information

The dataset was trained using the modified Yolov5 PyTorch model. The model was trained for 150 epochs that took 3.41 hours to complete.

Precision measures the percentage of correct predictions.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

TP = True Positives (Predicted as the occurrence of some object and it was correct)

FP = False Positives (Predicted as the occurrence of some object, but it was incorrect)

The precision for the vegetation under saturation augmentation was 75 percent and for missing tie was around 97 percent. The precision for the broken tie was around 20 percent and the reason for that was due to the insufficient dataset for the broken tie.

Figure 3: Saturation result

4.2 Exposure

4.2.1 Model Accuracy Information

The dataset was trained on the modified Yolov5 PyTorch model. The model was trained for 100 epochs that took 1.62 hours to complete.

The precision for the vegetation under exposure augmentation was 70 percent and for missing tie was around 93 percent. The precision for the broken tie was nearly to 0.

Figure 4: Saturation result

Figure 5: Exposure result

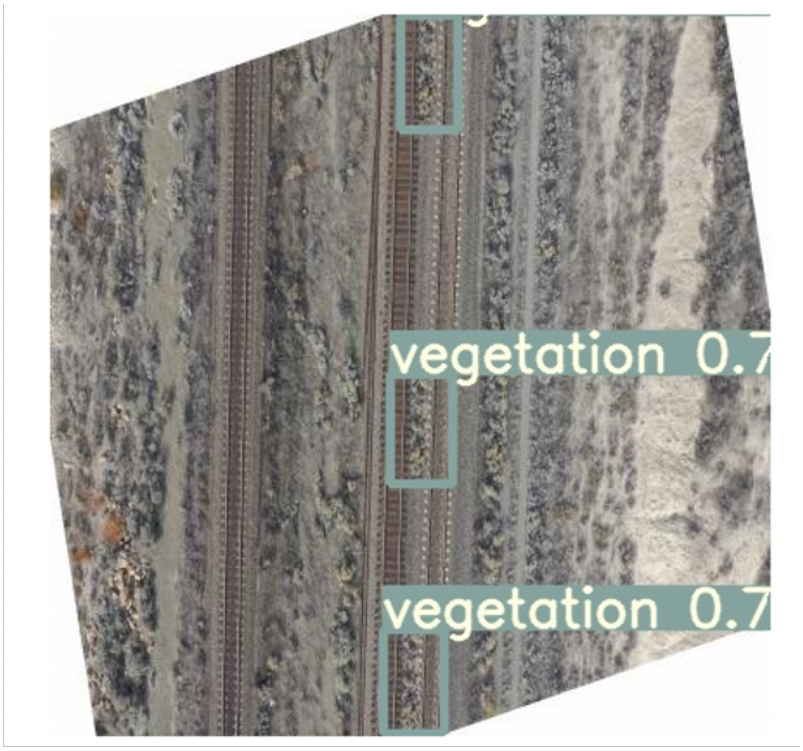

Figure 6: Exposure result

4.3 Mosaic

4.3.1 Model Accuracy Information

The dataset was trained on the modified Yolov5 PyTorch model. The model was trained for 100 epochs that took 3.50 hours to complete.

The precision for the vegetation under mosaic augmentation was very bad as it was nearly 0 percent and for missing tie was around 5 percent. The precision for the broken tie was nearly to 10 percent.

5 Conclusion

After comparing the 3 augmentation, it was concluded that the Saturation augmentation gave better result compared to Exposure and Mosaic. Also, after feeding the algorithm with pre-processed image and adding the vegetation and missing ties, we can conclude that the accuracy of various points of interest was directly proportional to the distance of the images. Images with broken ties were hard to identify from a greater distance. Also, the algorithm was quickly able to identify the vegetation and missing ties in the images from a closer distance. To summarize, we can say that it is recommended that the image is captured from a distance where it is easy for the algorithm to identify various points of interest, such as the vegetation, missing ties, and broken ties. For example, images are taken from 20m and 25m tend to give a higher accuracy rate than those taken from 50m and 70m. Please note that those are preliminary results and more data needs to be obtained.

Figure 7: Mosaic result

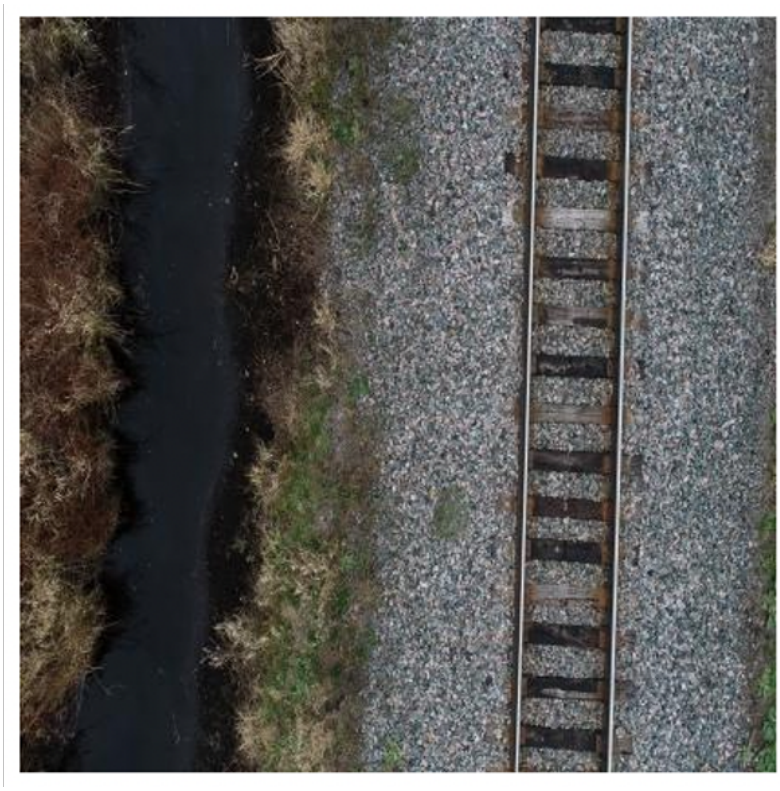

Figure 8: Mosaic result

6 Future Work

As a part of the future work, I am going to work on 4700 images where different types of augmentation will be tried and also number of epochs will be increase to see if it is affecting the accuracy.

References

[1] Bach Duy Khuat, Duong Thai Phung, Ha Thi Thu Pham, Anh Ngoc Bui, and Son Tung Ngo. 2021. Vietnamese Sign Language Detection Using Mediapipe. In *2021 10th International Conference on Software and Computer Applications (ICSCA 2021)*. Association for Computing Machinery, New York, NY, USA, 162–165. DOI: http://dx.doi.org/10.1145/3457784.3457810

[2] Abhiram Natarajan, Keshav Bharat, Guru Rajesh Kaustubh, Sai Praveen P. N., Minal Moharir, N. K. Srinath, and K. N. Subramanya. 2019. An Approach to Real Time Parking Management Using Computer Vision. In *Proceedings of the 2nd International Conference on Control and Computer Vision (ICCCV 2019)*. Association for Computing Machinery, New York, NY, USA, 18–22. DOI:http://dx.doi.org/10.1145/3341016.3341025

[3] Haijie Wang, Yanjie Ke, and Ge Yang. 2021. RT-LPDRnet: A Real-Time License Plate Detection and Recognition Network. In *2021 the 5th International Conference on Innovation in Artificial Intelligence (ICIAI 2021)*. Association for Computing Machinery, New York, NY, USA, 121–126. DOI:http://dx.doi.org/10.1145/3461353.3461391

[4] Ju Wu, Huan Shi, and Qinxue Wang. 2020. Contrabands Detection in X-Ray Screening Images Using YOLO Model *(CSAE 2020)*. Association for Computing Machinery, New York, NY, USA, Article 124, 5 pages. DOI: http://dx.doi.org/10.1145/3424978.3425106

[5] Tianhua Xie, Zebin Chen, Mingliang Cao, Pei Hu, Yuqing Zeng, and Zhigeng Pan. 2020. FACE DETECTION IN VR GAMES. In *2020 the 3rd International Conference on Control and Computer Vision (ICCCV'20)*. Association for Computing Machinery, New York, NY, USA, 7–10. https://doi.org/10.1145/3425577.3425579