Vision-based bicycle and motorcycle detection using a YOLO-based Network

C J M Dequito, I J L Dichaves, R J G Juan, M Y K T Minaga, J P Ilao1, M O Cordel II and N P A Del Gallego
College of Computer Studies, De La Salle University, Manila, Philippines

Email: joel.ilao@dlsu.edu.ph

Abstract. This paper describes a system that can distinguish between visually-similar objects, specifically bicycles and motorcycles, successfully from the vantage point of traffic surveillance cameras. The You Only Look Once (YOLO) is used as the main framework in this research due to its speed performance among various machine learning models and methods evaluated. We built a dataset consisting of motorcycle and bicycle images from different CCTV footage for this project. CCTV footage may vary on the angle of view from the object, image resolution, and ambient environment settings. Using this dataset, we trained YOLOv3-based models, and their performances were compared to the vanilla version of YOLOv3 and other pre-trained models. Four (4) models were trained and compared; the best-performing model is shown to be associated with a dataset with properly labeled data (i.e., marking every instance of the object of interest) and having the most number of instances in the training and testing set.

1. Introduction

The number of vehicles plying the roads, especially in urban areas, continues to grow. With it are associated problems such as heavy traffic, massive production of greenhouse gases, and an increased number of road accidents, to name a few. These burgeoning problems in road transportation paved the way for Intelligent Transport System (ITS) research, defined by [1] as "an integrated system of people, roads, vehicles designed to improve the road safety, efficiency, and comfort, as well as environmental conservation through the realization of smoother traffic congestion." One of the approaches that can be applied in ITS is data gathering and analysis. The data gathered can be performed with statistical analysis to acquire knowledge on traffic density in an area, air quality [11], or even detecting anomalies like accidents or traffic violations [12]. However, in the Philippine setting, roadways do not support this kind of technology yet [6].

While off-the-shelf detectors2 can be readily used to monitor vehicles on roads from the vantage point of traffic surveillance cameras, these systems do not perform well when detecting visually-similar vehicles such as motorcycles and bicycles. Relevant progress in the field includes the development of motorcycle and bicycle detection techniques such as using Gabor filters to detect pedaling movement to differentiate a motorcycle from a bicycle [9], ellipse approximation of bicycle tires [4], multi-feature extraction of 2-wheeled vehicles [10], and using extended Kalman filters to detect deformable parts in a bicycle (e.g., tires, seat, steering handle) [2]. However, most [4][9] of these developments

2 See https://github.com/amdegroot/ssd.pytorch
require that the two-wheeled vehicle be captured from a horizontal perspective, i.e., two wheels are seen in the frame.

From this study, we were able to produce a dataset of annotated images that cover motorcycles and bicycles. We have also produced a detector that focuses on detecting these types of vehicles which may be used by other systems.

2. Related Work
Various approaches in bicycle and motorcycle detection involve image detection and machine learning techniques. Takahashi and his colleagues [9] formulated a method to detect and classify bicycles against other two-wheeled vehicles on roads (e.g., motorcycles) using the pedaling movement present when using a bike. Spatiotemporal 3D Gabor filters were used to pinpoint the moving object regions in the extracted images. Afterward, the Histogram of Oriented Gradients (HOG) was calculated which was then fed to a Support Vector Machine (SVM) classifier to determine if it is a two-wheeled vehicle or a human. Overall, this approach attained an average accuracy rate of 76%. Although the formulated method can detect bicycles, it can only do so if the two-wheeled vehicle detected in the video was captured in a horizontal orientation. In addition to this, [4] proposed a method to detect bicycles by detecting tires using the convergent floating threshold method, and ellipse approximation. The algorithm fails when a person's foot is included in the frame of the bicycle's tire, or incorrectly detecting ellipse-shaped objects as tires. Zhang and Ling [10] on the other hand, extracted multiple features before classifying them.

Espinosa, Velastin, and Branch [3] developed an architecture based on faster R-CNN, which was trained and tested using a dataset composed of [5] and additional self-gathered data. The system attained a 75% mAP on images with occluded motorcycles and a 92% mAP on images with non-occluded motorcycles. The training process of the system from [3] involves first training the Region Proposal Network (RPN), and the detection network while minimizing the loss, and obtaining weights and biases. Finally, they combined the results of the pre-trained networks, which fused the parameters of a single network for detection and classification. They used the Stochastic Gradient Descent with Momentum for optimization of the training algorithm. The study from [3] also focuses only on motorcycle objects present in a frame as compared to our study which focuses on both bicycles and motorcycles. Furthermore, they used a more powerful GPU for training the Faster R-CNN based model as compared to this study. Table 1 contains a summary of existing works on vision-based bicycle and motorcycle detection.

Table 1. Summary of related works on 2-wheeled vehicle detection.

| Year & Authors | Framework | Dataset | # of Samples | Performance Results |
|---------------|-----------|---------|--------------|---------------------|
| Takahashi, Kuriya, & Morie, 2010 [9] | Spatiotemporal Gabor Filter, Histogram of Oriented Gradients (HOG), Support Vector Machines (SVM) | Self-gathered video footage | 100 images | 76% avg. accuracy rate |
| Fujimoto & Hayashi, 2013 [4] | Convergent floating thresholding, Ellipse Approximation | Self-gathered video footage | 30 images | 94.69% highest accuracy |
| Zhang & Ling, 2017 [10] | HOG, Cascade classifiers | Self-gathered video footage | 75 images | 96.96% highest accuracy |
| Espinosa, Velastin & Branch, 2018 [3] | Faster R-CNN | KITTI + self-gathered video footage | 7500 self-gathered images | 75% mAP occluded, 92% mAP non-occluded motorcycles |

3. Methodology
We model the bicycle and motorcycle detection problem as a binary classification task. For object detection, we used a pre-trained YOLOv3 model and performed transfer learning by using data...
gathered from online sources such as open-source datasets, and real-world traffic footage. After a model has been built, it is evaluated by comparing the results to the base accuracy of the algorithm. Changes such as establishing training and testing split ratio and adding more images to balance the dataset for both classifications were done for improvement.

3.1. Data Gathering

Saved CCTV footages of public highways were collected containing various perspective shots of passing vehicles. We also gathered videos and images from online sources such as Youtube and the Open Images Dataset. These videos serve as sources of cropped instances of motorcycles and bicycles used to build the annotated dataset for training the object detectors. From the data gathering, a dataset consisting of 6,671 and 4,590 annotated instances of motorcycles and bicycles, respectively was constructed under six categories: (1) daytime generic, a dataset of daytime images downloaded from the Open Images Dataset, (2) daytime highways (3) daytime inner streets (4) daytime avenue (5) nighttime inner streets, (6) rain dataset. Images that have visually-challenging environments such as night-time and rain were also collected.

Two (2) persons annotated the dataset: one person annotated the instances of bicycles and motorcycles in a frame while another person double-checked the first person's annotations. The criteria for acceptance include the distance of the motorcycle to the CCTV camera, the area covered by the bounding box, and the image's quality, especially in challenging scenarios.

Dataset augmentation techniques were also applied to expand the dataset. These include flip, rotate, brighten, and blur. Below is Figure 1 that demonstrates the difference of samples before and after data augmentation.

![Figure 1. A visualization of the dataset before and after data augmentation.](image)

3.2. Model Training

As shown in Figure 2, a pre-trained YOLOv3 model [8] is further trained using the collected training dataset to create a motorcycle and bicycle detector. Multiple models with various configurations were also created to determine the factors that correspond to better detection performance; these configurations include different data augmentation techniques and altering the split ratios of the training and test datasets. Generally, a training and test split of 70-30 was applied when training models 2 to 4, and no training and test split was applied to the first model.

---

3 From the Metropolitan Manila Development Authority (MMDA), the Pasig City Government, and De La Salle University-Manila for the road sections along EDSA, Shaw Blvd Obando Drive, and Taft Avenue, respectively.

4 Open Images Dataset, a dataset of ~9M images annotated with image labels, bounding boxes, etc., [https://storage.googleapis.com/openimages/web/index.html](https://storage.googleapis.com/openimages/web/index.html)
In addition, hyperparameters of YOLOv3 were altered to perform better on the desired vehicle identification. YOLOv3 was modified such that it detects bicycles and motorcycles only. For each training iteration, 64 samples are used as input, with each sample resized to 416x416, 64 was the value used due to GPU memory constraint. Furthermore, We used a learning rate equal to 1e-3 and scaled it by 0.1 on the 3200th and 3600th iterations, with the feature weights frozen. There are only two objects that this research aims to detect. With this we set the classes parameter to 2. The training stops after the 4000th iteration because the iteration limit was solely based on the YOLOv3 max batch formula, which is the total number of classes multiplied by 2000. For this research, there are only two classes thus resulting in 4000. Similarly, the filter parameters after each of the last convolution YOLO layers were also altered based on the configuration of the vanilla YOLO. Originally, YOLOv3 had a filter value 255 with 80 classes or objects. The formula for this is the number of classes plus 5 multiplied by 3. As a result, we used a value of 21 for the filter parameter. Other parameters that were not mentioned are set at their default values for the training.

3.3. Model Testing Descriptions

The motorcycle and bicycle object detectors were evaluated using the performance metrics accuracy, precision, recall, and mAP performance metrics. Before training multiple models, the proponents evaluated the performance of the pre-trained YOLOv3 to serve as a baseline on the various scenarios. Below is the summary table of the attributes of the models namely their training and test split ratio and the changes applied before training.
### Table 2. Summary table of the attributes of the model pre-training.

| Model               | Training-Test Split Ratio | Changes Applied                                                                 |
|---------------------|---------------------------|---------------------------------------------------------------------------------|
| Vanilla            | N/A                       | Vanilla Version of YOLOv3                                                          |
| 1                   | N/A                       | N/A                                                                              |
| 2                   | Training: 70% Testing: 30%| Balanced the instances of motorcycles and bicycles in the dataset. Detected improper annotation of images were fixed. |
| 3                   | Training: 70% Testing: 30%| Data Augmentation was performed and fixed detected improper annotation of images. |
| 4                   | Training: 70% Testing: 30%| Applied oversampling and undersampling to balance the dataset.                    |

After each modification of dataset and training, the proponents garnered performance metrics mentioned by assessing the models performance per frame. Details such as true positives, true negatives, false positives, and false negatives were observed and were used in assessing the performance of the models.

### 4. Results and Discussion

This section contains a per model discussion of the models with regards to the description of tests discussed in the previous section.

For the 1st model, a split for training and testing was not employed since it served as a trial run. It was observed that the model struggles to detect motorcycles in a cluster, a bicycle is more likely to be classified as a motorcycle rather than the opposite, and the model classifies some pedestrians as motorcycles, which might be caused by the training data since the riders were included. Moving on, results show in the second model that some objects of interest are missed by the detector, caused by improper annotation and labeling of training data. To define, an incomplete annotation of objects of interest results from not labeling all the objects of interest in a given frame, as shown in Figure 3. Model 2 performed best on the morning scenario with a 50.08% mAP. On the other hand, it performed the worst in the afternoon scenario with a 9.70% mAP.

![Figure 3. Complete annotation leading to identification of all instances (left) as compared to incomplete annotation leading to identification of only one instance (right).](image)

---

5th International Conference on Robotics and Machine Vision (ICRMV) 2021
Journal of Physics: Conference Series 1922 (2021) 012003
doi:10.1088/1742-6596/1922/1/012003
To improve the performance of model 1 and model 2, the improper annotations of the dataset images were fixed and data augmentation was applied. Initially, the dataset composition for this model was composed of 4,138 motorcycles and 2,183 bicycles. Data augmentation techniques, namely flip, rotate, blur, and brighten, were applied to the dataset to increase the number of instances for the model. Model 3 performed best in the noon scenario with an mAP of 65.43%. Meanwhile, it performed the worst in the night scenario with an mAP of 31.69%.

The 4th model was made to check the performance of the model given a balanced dataset. In the previous model, more motorcycles are present both in the Pasig and Rain categories of the dataset. To balance the dataset, oversampling and undersampling was utilized. Frames with an abundance of motorcycles compared to bicycles were removed to keep a close balance of the number of instances between the number of motorcycles and bicycles. The number of images was also reduced to 1500 with a +/-100 threshold of allowable shortage or excess of images. This method was also applied to reduce the bias of the model in identifying more motorcycles than bicycles. Since the ratio of motorcycles and bicycles in this model was increased, especially the instances of bicycles, this model performed more bicycle identifications than any other trained model. The 4th model performed best on the noon scenario with a 70.67% mAP, and performed the worst on the rainy scenario with a 33.71% mAP.

5. Analysis of Model Performance
All models were tested in five scenarios, namely morning, noon, afternoon, night, and rain. These scenarios were chosen for the tests because they generally account for the variations in terms of ambient lighting conditions for the gathered video datasets. This section will include observations in the performance of the models in the said scenarios and speculations behind the behavior of the model.

5.1. Morning
The morning scenario resulted in the best performance for all the trained models due to good lighting and camera placement, as evident in Figure 3. In this scenario, the videos were taken by a camera that can pan, giving a diverse set of perspective shots for the detector. A notable observation in this scenario is the models’ ability to identify motorcycles given a partial part of the object. Although the performance of the models in this scenario is said to be the best among all scenarios, there are still cases that the model can improve on. For instance the detector identifies parts of a car to be a motorcycle (left) and having two detections at one object (right) as seen in Figure 4. The alteration of the dataset, specifically adding more instances with morning conditions and balancing the instance count may have helped in the improvement of the model.

Figure 4. The model detecting a partial part of a vehicle as a motorcycle (left) and the model identifying two bicycles in one instance (right).
5.2. Noon
The configurations of the video are the same with the morning scenario since the place where the video was taken and camera that was used were the same.

Observations in this scenario include: the models’ capability in detecting objects far from the camera (1), identifying an object of interest as both a bicycle and motorcycle (2), and identifying partially occluded motorcycles as a single motorcycle (3). This scenario also reflected the model having a maximum in the number of objects detected per frame, resulting in several false negatives, as shown in Figure 5 (left). This identification error sprouted from the improper annotation of the dataset as discussed in the previous section. Lastly, a common error in detection for the models tested in this scenario would be: a pedestrian, a puddle, and motorcycle-like objects like e-scooters as seen in Figure 5 (right).

![Figure 5](image1.png)

**Figure 5.** A scenario where the model only detect limited objects in a frame (left) and the model misidentifying a puddle, a pedestrian, and an e-scooter in a single frame.

5.3. Afternoon
The afternoon scenario that was used to test the models was taken from the Pasig City video collection, which has the highest camera placements compared to all other videos. Because of the traffic surveillance camera’s placement, two-wheeled vehicles are typically seen in mostly top view; thus, the models performed lowest in this scenario in comparison to other scenarios. The system identifies pedestrians more frequently than bicycles in this scenario. In this scenario, the model also has the tendency to group an instance of a motorcycle and bicycle as one instance of a motorcycle. Lastly, the presence of a tricycle in this video is identifiable but it is not limited to the motorcycle part of the vehicle. These observations are shown in Figure 6.

![Figure 6](image2.png)

**Figure 6.** The model identifying a pedestrian as a bicycle (left), the model detecting a bicycle and a motorcycle as a motorcycle (center), and the model detecting a tricycle as a motorcycle (right).

5.4. Night
The specifics such as the area where the footage was taken and the camera angle of the night-time scenario video are the same with the morning and noon scenarios. Although the video was taken with
poor lighting, the detector's performance is not as bad as when the detector was run in the afternoon scenario. There are also notable models' performance observations specific to this scenario, such as the ability to detect instances even at high speed where instances may be blurry. The most prominent detection error present in this scenario is the detector's inability to detect occluded instances as separate instances. These observations are demonstrated in Figure 7.

![Figure 7](image)

**Figure 7.** The model detecting a vehicle in high speed (left) and the model identifying two bicycles as one instance (right).

5.5. Rainy

The rainy scenario for testing the models is notably different from other videos used in for performance testing. Aside from the location being different, water droplets are present in the camera, distorting a region in the frame and making it difficult for the detector to identify instances of objects of interest. Because of the rainy weather, bicycles are not found in the video.

The problem of inconsistent detection still persists since object tracking is not a capability of the models. To add to this, the scenario of detecting clusters of objects of interest is also present in this scenario. This problem is however solved as the objects of interest appear larger within the frame of an image as they approach the camera. The system frequently identified a pedestrian on the left side as a motorcycle on all models, as seen in Figure 8. In addition to a pedestrian sharing similar features to a motorcycle: the torso and head, the blurred part of the frame due to moisture also adds to the system's confusion.

![Figure 8](image)

**Figure 8.** Pedestrian identified as motorcycle in the rainy scenario video.
The model best performed in the rainy scenario since the given camera angle is suitable for the model although the vehicles may not be clearly seen due to the circumstances of the scenario. On the contrary, the model had the worst performance on the afternoon scenario. The given camera angle of the scenario in addition to the lighting during the time of the video may have affected the performance of the model in this scenario. A summary table of the performance of the model in bicycle and motorcycle detection is listed on the tables below followed by the discussion of the performance of the model on each scenario.

| Table 3. Summary of Each Model’s Motorcycle Performance on Various Scenarios. |
|---|
| Model | Metric | Ambient Condition |
| | | Morning | Noon | Afternoon | Night | Rainy |
| 1 | Accuracy | 58.13% | 44.66% | 17.88% | 24.95% | 53.06% |
| | Precision | 67.68% | 67.67% | 96.94% | 77.25% | 77.40% |
| | Recall | 56.15% | 64.70% | 17.76% | 25.66% | 62.79% |
| 2 | Accuracy | 30.38% | 24.70% | 5.47% | 53.06% | 38.61% |
| | Precision | 89.74% | 66.66% | 88.13% | 77.40% | 86.88% |
| | Recall | 23.48% | 28.06% | 3.44% | 62.79% | 41.00% |
| 3 | Accuracy | 63.78% | 67.34% | 43.04% | 48.89% | 68.26% |
| | Precision | 66.66% | 84.23% | 93.37% | 87.84% | 86.93% |
| | Recall | 40.86% | 72.47% | 42.70% | 33.70% | 76.07% |
| 4 | Accuracy | 70.49% | 65.49% | 30.39% | 37.88% | -- |
| | Precision | 100.00% | 94.28% | 0.89% | 100% | -- |
| | Recall | 22.75% | 5.86% | 2.11% | 3.35% | -- |

| Table 4. Summary of Each Model’s Bicycle Performance on Various Scenarios. |
|---|
| Model | Metric | Ambient Condition |
| | | Morning | Noon | Afternoon | Night | Rainy |
| 1 | Accuracy | 40.67% | 65.49% | 30.39% | 37.88% | -- |
| | Precision | 100.00% | 94.28% | 0.89% | 100% | -- |
| | Recall | 22.75% | 5.86% | 2.11% | 3.35% | -- |
| 2 | Accuracy | 16.88% | 42.02% | 12.50% | 12.85% | -- |
| | Precision | 100.00% | 100.00% | 6.96% | 100.00% | -- |
| | Recall | 5.11% | 0.36% | 17.69% | 0.88% | -- |
| 3 | Accuracy | 56.99% | 77.06% | 61.42% | 63.93% | -- |
| | Precision | 97.87% | 90.95% | 14.58% | 100.00% | -- |
| | Recall | 48.93% | 33.83% | 19.57% | 11.00% | -- |
| 4 | Accuracy | 54.98% | 81.78% | 69.38% | 60.52% | -- |
| | Precision | 96.96% | 65.11% | 33.33% | 87.25% | -- |
| | Recall | 42.66% | 59.95% | 21.28% | 15.47% | -- |

6. Conclusion
In this study, we trained an object detector by retraining the vanilla version of YOLOv3 with different configurations. The resulting model better detects two-wheeled vehicles in videos captured from a roadside traffic camera perspective. In this project, a dataset was built, covering different camera perspectives and ambient conditions and taken from various sources. The performance of the model can be attributed to the variety of images with different angles. Moreover, the annotation of the dataset is another factor that helped the detector perform well; better image annotation yields a better performance for the model. Lastly, dataset augmentation was used to obtain more samples for training the different object detector models.

The results of this study can be extended to other similar studies or projects. First, the dataset can be used to train other models that are specific to object detection. The model can also perform in different
scenarios with different environmental factors. More work could be done to improve the performance of the model. Aside from performing more data augmentation techniques, having more examples of motorcycles and bicycles the models can learn from can improve its performance. In line with this, the addition of samples from similar datasets that deal with object detection, such as MIOvision Traffic Camera Dataset (MIO-TCD) [7] can be used. In addition, possible future works could include the additional usage of reinforcement learning to the training process to help the model perform better by adding in object tracking for consistent detection of objects of interest. Finally, using a later version of available models such as YOLOv4 can also improve the performance of the model.

7. REFERENCES

[1] Andersen J and Sutcliffe S 2000 Intelligent Transport Systems (ITS) - an overview. Proc. DECOM-TT 2000: IFAC Conference on Technology Transfer in Developing Countries: Automation in Infrastructure Creation (Pretoria, South Africa, 5-7 July 2000) 33 18 pp 99 - 106

[2] Cho H, Rybski P E and Zhang W 2010 Vision-based bicycle detection and tracking using a deformable part model and an EKF algorithm Proc. 13th International IEEE Conference on Intelligent Transportation Systems (Funchal, Madeira Island, Portugal) pp 1875-80

[3] Espinosa J E, Velastin S A and Branch J W 2018 Motorcycle detection and classification in urban scenarios using a model based on fasterR-CNN Proc. 9th International Conference on Pattern Recognition Systems (Valparaiso, Chile, 22-24 May 2018) pp 91-6

[4] Fujimoto Y and Hayashi J 2013 A method for bicycle detection using ellipse approximation Proc. The 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision (Incheon, South Korea, 30 January – 1 February 2013) pp 254-7

[5] Geiger A, Lenz P, Stiller C and Urtasun R 2013 Vision meets robotics: The KITTI Dataset Int. J. Rob. Res., 32 11 pp 1231-7

[6] Ilao J, Dang M, Daquitoa A, Ramos T and See F 2007 Hypothesis generation of vision-based vehicle tracking system. Proc. 7th Philippine Computing Science Congress (Boracay, Philippines, 23-24 February 2007)

[7] Luo Z, Branchaud-Charron F, Lemaire C, Konrad J, Li S, Mishra A, Achkar A, Justin E and Jodoin P 2018 MIO-TCD: A new benchmark dataset for vehicle classification and localization IEEE Transactions on Image Processing 27 10 pp 5129-41

[8] Redmon J and Farhadi A 2018 Yolov3: an incremental improvement CoRR,abs/1804.02767. Retrieved from http://arxiv.org/abs/1804.02767

[9] Takahashi K, Kuriya Y and Morie T (2010, Nov). Bicycle detection using pedaling movement by spatiotemporal Gabor filtering. Proc. TENCON 2010: 2010 IEEE Region 10 Conference (Fukuoka, Japan) pp 918-22

[10] Zhang Y and Ling Q 2017 Bicycle detection based on multi-feature and multi-frame fusion in low-resolution traffic videos arXiv e-prints

[11] Diaz G, Macià H, Valero V, Boubeta-Puig J and Cuartero F 2020 An Intelligent Transportation System to control air pollution and road traffic in cities integrating CEP and Colored Petri Nets Neural Computing and Applications 32 10.1007/s00521-018-3850-1

[12] Wong S F, Mak H C, Ku C H and Ho W I 2017 Developing advanced traffic violation detection system with RFID technology for smart city Proc IEEM: EEE International

See https://github.com/Tianxiaomo/pytorch-YOLOv4
8. ACKNOWLEDGEMENTS

We would like to acknowledge and thank the Department of Science and Technology (DOST) for funding “TITAN: Vision-based Traffic Information & Analysis” with project number 07595 that initiated this research.