Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Literature Review and

Statement of Research Intent

Project Number: 73

Identifying intruders on scooters entering carparks

Vinayak Joshi

Adwait Mane

Prof. Robert Amor

26/04/2024

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Vinayak Joshi

Name: Vinayak Joshi

ABSTRACT: This review examines the current methodologies in AI for real-time object detection, particularly for e-scooter detection, going over the current state of the literature, highlights the evolution of modern object detection algorithms from R-CNN to the almighty YOLO, explores the challenges of data scarcity, the need for more robust models and better performance metrics, as well as commonalities, conflicts, and holes across the literature.

Introduction

In the spring of 1997, Garry Kasparov, world champion and undefeated for over a decade, took a shot across his bow when he was defeated by IBM's chess engine, Deep Blue. Today, we look back to that moment as one of the first times that a machine stood triumphant against man, and although Deep Blue was only a symbolic representation of AI, the world is now full of what we call such. With the recent resurgence in the popularity of AI due to its massive advancements, the topic has become important as it becomes more integrated with our everyday lives.

The purpose of this literature review is to understand the current methodological practices employed by AI, not for playing chess – but for the detection of objects, where we will identify any of their gaps in the areas they are applied, and attempt to establish "next steps" for further research.

This literature review will follow a structured approach, beginning with a brief overview of foundational concepts and methodologies that underpin all machine learning models. We will then narrow our focus to studies relevant to real-time object detection, analyzing methodological approaches and recent advancements. By addressing these aspects, we aim to provide a comprehensive overview of the current state of object detection and its applications in real-time object detection.

Finally, the importance of this review, particularly regarding the case study, lies within the fact that real-time AI detection overlaps heavily with applications regarding security, such as the prevention of theft mentioned above, which falls under the category of a safety-critical system. Such systems are crucial to prevent extreme damage to life or property and protect valuable assets. There are many things yet to iron out in the realm of AI that pertain to such systems, and hence, a focus on current standard practices, their pitfalls, and potential improvements becomes paramount.

1. Literature Review

3.1 An overview of object detection

Just as the airplane was inspired by birds, the bullet train from the kingfishers, or velcro from the burdocks, so too was the neural network from the human brain. This, in fact, is one of the core engines powering many modern AI systems, including that which is concerned by this review for the purpose of object detection.

A neural network is exactly what it sounds like – a network of neurons. Each neuron holds a numerical value, which passes another value to all other neurons connected to it in a forward direction. This other value is nothing more than the output of a function, where the input is the previous neuron's held value.

Of course, these networks can incorporate many layers, and can span into the billions of neurons with many billions of connections. Generally speaking, the input of a neural network in the specific domain of object detection is nothing more than the entire grid of pixels that make up the image in which we wish to find the object. These pixels are oftentimes flattened from their normal two-dimensional arrangement into a single-dimensional array. While this is the soul of many modern techniques used in the field of object detection, there are many different forms of neural networks, and many different algorithms to leverage them for object detection (such as YOLO) – their specific structure tweaked for the problem at hand.

3.2 Current state of literature

Given that e-scooters have only risen in popularity relatively recently, the first boom of modern rentals appearing sometime in 2017 with the inception of companies such as Bird and Lime, there is incredibly limited literature on the current topic. Any current literature has come about due to some of the problems e-scooters cause, such as conflict with pedestrians [9][10][13] or accidents and injuries [17]. It just so happens that e-scooter detection with machine learning is a potentially valuable way of gathering information and statistics on riders to combat these problems. In fact, this very literature review also spawned out of a need for better e-scooter detection techniques given their use in burglary attempts. Of these pieces of literature, some are not focused on e-scooter detection as a whole but rather on specific subproblems, such as the detection of occluded e-scooter riders or those that have fallen onto the ground. The advancement of this topic remains crucial given its bleak state, and to combat current serious issues that e-scooters can cause if misused.

3.4 Modern object detection methods

From the rigid and clunky template-matching techniques of the 1970s [1] to modern neural networks, object detection has come a long way. The advent of the neural network was a diamond moment in the AI landscape. Once NVIDIA smashed through the door with their CUDA framework, allowing anyone to train their models with unrivaled speed [18], researchers could work tenfold to advance the field.

The rise of the Convolutional Neural Network (CNN), inspired once again by the human brain, allowed machines to, at last, generalize the qualities and characteristics of the elements of their input rather than finding an exact or near-exact match. From this, many object-detection algorithms were born that utilized a CNN, including R-CNN, Fast R-CNN, Faster R-CNN, and the almighty YOLO,

R-CNN, or “Region-based Convolutional Neural Network”, employed a technique called “selective search” that would first divide the image into different regions where an object may lie, extract the features of these regions, then send them off to the CNN for individual analysis [19]. Although this improved detection accuracy quite a bit, it was rather slow and not suited for real-time use.

Fast R-CNN, just as the name suggests, was meant to address the issues of normal R-CNN. The difference this time is that rather than splitting the image into regions before sending them through the CNN, the image is simply sent through intact *and then* split into regions, after generating a feature map. This saved the computation time of sending multiple regions through a CNN but still required analysis of the different regions anyway. More speed was still highly desired.

Faster R-CNN, following the highly original and exciting naming convention, improved upon Fast R-CNN by introducing something called a “Region Proposal Network” or “RPN”. This RPN is a convolutional neural network of its own, and rather than having a separate step to split the image into different regions, the RPN did it implicitly during the detection process, saving more computational time [20].

Following Faster R-CNN came the almighty YOLO, which unfortunately broke the exciting and original naming scheme but improved performance drastically by simply performing image sectoring, feature extraction, and prediction within a single forward pass [8], hence the name “You Only Look Once”. This, however, came at the expense of some accuracy.

3.5 Common findings – challenges, methods, and assumptions

As it turns out, the YOLO algorithm has become the de facto choice for object detection [6], especially in a real-time context where speed is of utmost importance and false positives can be tolerated, which remains true even for e-scooter detection. While at this point models are good at generalizing in their narrow domain, an immense amount of quality training data is needed in order to create any good results, and this is one of the significant challenges that seem to underpin most real-time detection applications, especially within e-scooter detection [8][9]. The reason for this is that it’s simply difficult to capture the large amount of data required to have an effective model while also generalizing the conditions such that the object in question is isolated from the surrounding environment, especially given that e-scooter riders are oftentimes found in very complex, dynamic environments such as bustling cities. Interestingly enough, some clever techniques have been used to help data collection by leveraging models that are already great for detecting *pedestrians*, and then simply enlarging the bounding box to include the e-scooter if there is one [9][10] and classifying those, after first collecting video footage of a busy section of a city. This seems to be done in “filtering” steps, where YOLOv3 is generally used for detecting pedestrians, the bounding box enlarged, and then MobileNetV2 is used to detect if any of these boxes contain e-scooters. There is an implicit assumption here that e-scooters will always have riders, although to be fair there are hardly any real-world examples of requiring the detection of e-scooters without a rider. For the time being, this at least indicates that the detection of e-scooter riders is perhaps an extension of the problem of detecting pedestrians – or even *human* detection.

Despite these clever techniques for data gathering. the difficulty is worsened by the fact that many a time, partial occlusion of objects in dynamic environments (such as the outdoors) renders the detection of the desired object in question incredibly inconsistent, necessitating the development of entirely new strategies to combat this as well as requiring even more data in an attempt to generalize the features of objects even more [9]. What’s more, many of the training datasets used are simply borrowed from third parties, such as COCO [11][12] or simply Google Images [17], with an implicit assumption that these are appropriate for all environments [8].

Alongside the general challenges with data collection also comes speed. The larger the model, the slower it is given the extra time it takes to process a new image input. The few sources available in the realm of e-scooter detection make almost no explicit mention of the suitability of their proposed methodological fixes for feasibility in real-time detection. Instead, they only focus on accuracy, true positives, false positives, true negatives, and false negatives. This is important because a theoretically excellent model *does not* equate to a practically excellent model.

A final interesting point is that the literature seems to *acknowledge* that e-scooter riders exhibit different characteristics of movement compared to other things [9], such as vehicles or pedestrians. This could indicate that, much like how the detection of e-scooter riders is perhaps an extension of the problem of detecting pedestrians/humans mentioned before, the problem may also be an extension of motion analysis or even pose detection (as riders have a different pose when riding scooters as compared to walking), spawning a new research question on how such techniques used in these sub-fields could possibly be leveraged to improve the detection performance of e-scooters. A potential place to start looking is how current state-of-the-art models detect cyclists, motorcyclists, moped riders, or something similar to these.

3.6. Common findings – differing results and research holes

While YOLO seems to be used willy-nilly across the board, with proposed improvements by using the MobileNetV2 CNN architecture [9][10], an article published by California Polytechnic State University [12], aiming to count the number of e-scooter riders passing by in traffic for the purpose of Santa Monica City Council to investigate pedestrian conflicts had instead made use of RetinaNet [13], with great results shown in their demo. Interestingly enough, it seems the primary issue with advancing the current state of the literature is the plethora of different environments, use cases, and constraints of e-scooter detection. Some are in environments where there is occlusion, some are not. Some detect e-scooters from overhead, some straight on. Some attempt to detect e-scooters going very fast and some do not. Some attempt to classify e-scooters from a live broadcast and some do not. In truth, the underlying patterns of the current literature indicate that techniques for improvement are for specific sub-problems of machine learning *in general* (e.g. detection for occluded objects), and the fact that e-scooters are the primary focus doesn't really matter; some techniques (such as leveraging pedestrian models and enlarging the bounding box) are indeed clever, but somewhat agnostic of the object in question. The same technique could be applied to help the detection of skateboarders or rollerbladers, for instance.

Ignoring metrics beyond accuracy and precision is almost universal in all sources, where they simply do not appear to care about the power efficiency of running the models (with the exception of Santa Monica's implementation [13]), the ease of their deployment and integration, versatility, or robustness. It is fundamentally the case that these models are meant to be used in the real world, and ignoring such metrics is tunnel-visioned at best, and careless at worst. The question becomes why this is the case, as even metrics such as consistency are hardly used, indicating a huge hole in current performance evaluation procedures [14][16]. It is also the case that there are simply no standard benchmarking and evaluation guidelines [15], and this leads to ambiguous and sometimes meaningless comparisons. It becomes disingenuous to attempt to compare the results of different pieces of literature, as they are used in different applications, and environments, have different training and testing datasets, use different model architectures, and are sometimes concerned with different problems entirely. This isn’t the fault of the authors but reflects the current void in research for this particular area.

3.7. Conclusion

Although quite a niche topic, this literature review has revealed a promising future for real-time object detection of e-scooter riders as a means to prevent some of the issues that their misuse causes, such as burglary, traffic accidents, conflicts with pedestrians, and more. There are some notable commonalities with the current literature that can help future researchers guide their efforts, such as potentially delving into pose detection, movement evaluation, and/or leveraging current pedestrian detection to see what sticks. However, the lack of standardized benchmarking protocols, the need for more comprehensive evaluation metrics, and specific sub-issues dragging the focus away from e-scooter detection in particular such as occlusion-aware detection methods or fallen riders all necessitate further research into this area. Although the current state of the literature on this topic is bleak, the severity of the problems remains no smaller than in other areas of the world that have far more research, such as motorist/cyclist accidents, or general pedestrian safety.

4. Project Scope

4.1 Research Intent

4.1.1 Understanding integration challenges

There comes a time in every engineer's life when they come up with a grand idea that simply bends the Iron Triangle. It is therefore absolutely necessary that we look into integration challenges, such as cost, as just because a technique is theoretically excellent does not mean it is practically excellent. Balancing the performance of the actual model with practical constraints is an absolute necessity. The research question for this will be “How can we improve current model performance while also balancing them with real-world constraints?”.

4.1.2. Assessment of performance metrics

There are many performance metrics that are used for the assessment of AI tools. A question that is severely underlooked is what these performance metrics exactly are, whether or not they are good, and if they’re worthy of being called a performance metric in the first place. Companies especially will boast arbitrary numbers on obscure benchmarks and a similar sentiment can oftentimes be found in research papers, which we will critically evaluate.

4.1.3. Addressing data scarcity and quality

As discovered in the literature review, access to large, high-quality datasets is incredibly difficult to get one's hands on. Although some clever tricks have been used to gather good data, we aim to research possible techniques, methods, or tricks to generate high-quality data without much effort. Ideally, this would be done in a completely hands-free automatic fashion. The research question for this will be “How can we make data collection/data generation easier, while also maintaining a high level of quality?”.

4.1.4. Enhancing model robustness

Real-time object detection demands high robustness against varying environmental conditions. Our research will also focus on finding techniques to enhance model generalization, and see if there is anything about e-scooters in particular that could be leveraged to enhance detection accuracy and performance. As mentioned previously, this could include diving into pose detection, movement analysis, or checking for patterns with other detection techniques for “similar” objects. The research questions for this will be “How can we ensure that e-scooter models are able to work in a variety of environments, and how can we leverage other areas of literature (e.g. pose detection) to elevate model performance?”

4.2. Gantt Chart

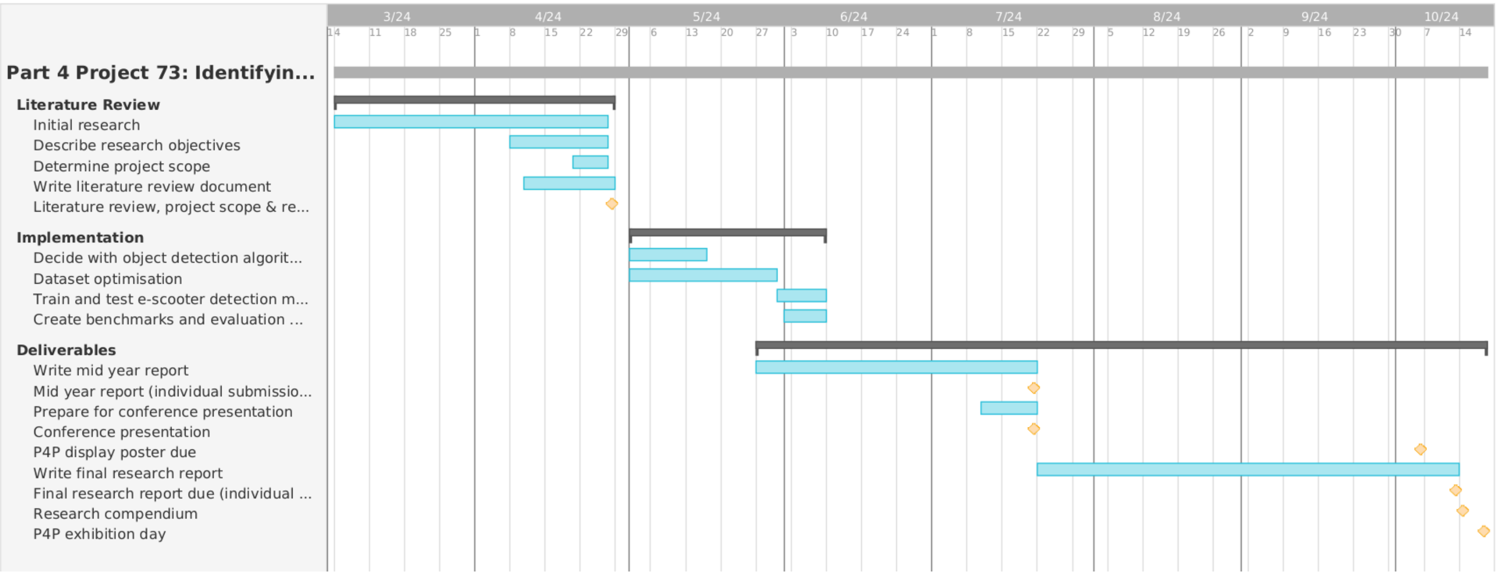

Fig. 1. Timeline of the research project

4.3. Project Boundaries

We do *not* aim to make a fully functional, ready-to-deploy model suitable for real-world use. Rather, we aim to make a prototype model implementing the adjustments we see fit from both best practices and our research findings. We also do not aim to conduct real-world, end-to-end testing, as the resources required for this are out of our reach. Instead, we will rely on testing and training splits to evaluate model performance. Given this fact, ease of deployability, and other practical considerations we will make heuristic evaluations rather than making an actual deployment. We also do not aim to find a pot of gold laying about on the internet in terms of a quality dataset but rather come up with methods on how to make quality data collection easier as well as, ideally, *generate* quality data autonomously.

References

[1] N. S. Hashemi, R. B. Aghdam, A. S. B. Ghiasi, and P. Fatemi, "Template Matching Advances and Applications in Image Analysis," arXiv:1610.07231.pdf, 2016.

[6] M. Hussain, "YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection," Machines, vol. 11, no. 7, p. 677, Jul. 2023. doi: 10.3390/machines11070677. [Online]. Available: https://www.mdpi.com/2075-1702/11/7/677

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[9] M. Serrano, P. Miró, and S. D'Antonio, "E-Scooter Rider detection and classification in dense urban environments," in Sensors, vol. 22, no. 23, 2022

[10] X. Zhou, W. Wang, T. Zhu, J. Ye, and T. Lin, "Detection of E-scooter Riders in Naturalistic Scenes," arXiv:2111.14060, 2021

[11] T. Lin et al., "The COCO Dataset," cocodataset.org, 2014. [Online]. Available: https://cocodataset.org/#home

[12] T. Lin et al., "Microsoft COCO: Common Objects in Context," in European Conference on Computer Vision (ECCV), 2014

[13] X. Li, "E-Scooter Counting on Sidewalks with Machine Learning," dxhub.calpoly.edu, 2022. [Online]. Available: https://dxhub.calpoly.edu/challenges/escooter-counting-on-sidewalks/

[14] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," arXiv:1708.02002, 2017

[15] A. Mhatre, S. Kabir, M. Donahue, and S. Tulsiani, "Why Accuracy Is Not Enough: The Need for Consistency in Object Detection," arXiv:2207.13890, 2022

[16] U. Shankar, V. Jain, and D. Gupta, "A Comprehensive Study of Real-Time Object Detection Network Across Multiple Domains: A Survey," arXiv:2208.10895, 2022

[17] N. T. Hieu, M. Q. Minh, and N. M. Quan, "Electric Scooter and Its Rider Detection Framework Based on Deep Learning for Supporting Scooter-Related Injury Emergency Services," in [Proceedings of the International Conference on System Science and Engineering (ICSSE)], Auckland, New Zealand, March 2021 doi: 10.1007/978-3-030-72073-5_18

[18] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with CUDA," ACM Queue, vol. 6, no. 2, pp. 40-53, Mar./Apr. 2008, doi: 10.1145/1365490.1365500.

[19] R. Girshick, “Fast R-CNN,” 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, Dec. 2015, doi: https://doi.org/10.1109/iccv.2015.169.

[20] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” arXiv.org, 2015. https://arxiv.org/abs/1506.01497