Combining YOLO and Visual Rhythm for Vehicle Counting

Victor Nascimento Ribeiro

Nina S. T. Hirata

Institute of Mathematics and Statistics - University of São Paulo

victor_nascimento@usp.br

Objectives

Counting vehicles is an important task for traffic control and management of transport infrastructure. This problem can be addressed employing methods for vehicle detection in videos. We consider scenarios where static cameras are used to capture videos from a top-view perspective of vehicles moving in one direction. Rather than detecting and tracking vehicles, we propose a combination of Visual Rhythm, a technique that generates time-spatial images from videos, and YOLO, a well known model for object detection. This combination enables us to selectively process frames containing relevant information, thereby leading to enhanced efficiency.

Materials and Methods

We define the vehicle counting problem as the task of quantifying the number of vehicles that cross a designated line (counting line) in the camera's field of view. This can be accomplished by detecting the vehicles, tracking them, and counting them when they cross the counting line in a frame-by-frame approach. To avoid processing every frame of the video, which is a computationally costly procedure, we use Visual Rhythm (VR) [1] to identify frames where an object overlaps the line, and then employ a vehicle detection model on that specific frame to recognize vehicles.

Let f be a video with T frames of size $M \times N$. The VR of f is built by extracting the row corresponding to the counting line of each of the T frames and stacking them along the time axis. This results in an image with size $T \times N$, where each row is a line from a frame, as shown in Figure 1. Note that, around frame 20 the vehicle crosses the line and therefore in the VR image there is a mark of the vehicle.

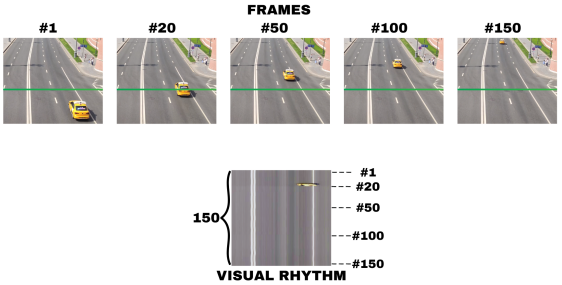

Figure 1: Visual Rhythm Generation

Both for mark and vehicle detection we use YOLO (You Only Look Once) [2], a state-of-the-art deep learning based real-time object detection model.

The main steps of the proposed method are illustrated in Figure 2. In the first step (a), we create a Visual Rhythm (VR) image for segments of T consecutive frames. In step (b), YOLO is employed to detect marks within the VR image. Subsequently, in step (c), we extract the corresponding frame for each detected mark. Next, we must certify that the mark in question indeed corresponds to a vehicle. In

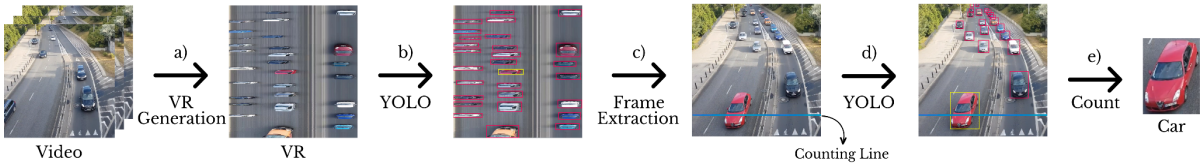

Figure 2: Data flow in the VR–based video counting vehicles

step (d), YOLO is used once more to detect all vehicles within the frame, and the vehicle that best matches the size and position of the mark is selected in step (e) to update the vehicle class count. We partition long video sequences into non-overlapping segments of length T to manage VR image size for YOLO processing.

Results

To evaluate the proposed method, we use 4 videos available on YouTube. Both vehicles and marks on VR were annotated, and separated on training, validation and testing sets, making sure that no same vehicle is present in more than one of these subsets. Then, we fine-tuned a pre-trained YOLOv8-small model for vehicle and mark detection. In our experiments, the counting line was fixed at height 120 of the frame and we used video segments with length $T = 900$ (30 seconds) to build the VR images. We compared our approach to Roboflow's frame-by-frame vehicle counting system [3], which utilizes ByteTrack for real-time object tracking. Table 1 shows the results on the test set of three videos. Both systems operated in identical environments with the same model weights. The proposed VR-based method is about three times faster than tracking-based, as it only performs vehicle detection on selected frames.

Table 1: Counting accuracy (%) in test set

System	Frame rate FPS	Video 2	Video 3	Video 4
VR	186	100	98.9	98.5
Bytetrack	56	100	99.0	98.4

In Table 2 we show recognition accuracy regarding vehicle class. Among the categories, 'Car' achieved the highest accuracy at 99.1%, whereas 'Motorbike' and 'Pickup' had lower accuracies of 81.6%, and 77.0%. We note that these accuracies refer to the set of detected vehicles on the test set.

Table 2: Classification accuracy (%) of VR approach in test set

Car	Bus	Motorbike	Pickup	Truck	Van
99.1	86.0	81.6	77.0	96.2	92.9

Conclusions

The proposed VR-based method for vehicle counting is approximately 3 times faster than a tracking-based method. This gain in speed is thanks to the efficient selection of only frames of interest, avoiding the need to perform vehicle detection on all frames of the video. Moreover, the counting accuracy is on par with the tracking-based method, indicating its efficacy. There is room for improving vehicle type recognition.

Acknowledgements

MCTI (Brazil), law 8.248, PPI-Softex - TIC13 - 01245.010222/2022-44, FAPESP 2015/22308-2.

References

[1] S. Guimarães, M. Couprie, N. Leite, and D. A. Araújo, “A method for cut detection based on visual rhythm,” in Proceedings XIV Brazilian Symposium on Computer Graphics and Image Processing, 2001, pp. 297–304.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” CVPR, pp. 779-788, 2016.

[3] B. Dwyer, J. Nelson, J. Solawetz et al., “Roboflow (version 1.0) [software],” https://roboflow.com, 2022, computer vision.