A Multiple Vehicle Tracking and Counting Method and its Realization on an Embedded System with a Surveillance Camera

Yi-Hsuan Hsu, Ssu-Yuan Chang and Jiun-In Guo
Department of Electronics Engineering and Institute of Electronics, National Chiao Tung University, ROC
E-mail :{piccolo1992514, jiguoccu}@gmail.com

***Abstract*—This paper proposes a tracking-by-detection method with a weighted scoring mechanism to associate the trackers and the detection results for accurate tracking and counting vehicles in a surveillance application. For the vehicle detection, the proposed method uses our robust PVA-lite deep learning model to detect vehicles. The experimental results show that the proposed method can achieve more than 95% in the average counting accuracy.**

I. INTRODUCTION

With the dramatically increasing of IOT applications, vehicle counting has become a very popular topic. To count the vehicle accurately, tracking the targets precisely is of great importance because it can avoid counting the same vehicle multiple times and provide the car flow estimation. Long-term visual tracking in unconstrained environments is critical for many applications such as video surveillance, Advanced Driver Assistance Systems (ADAS), home safety applications, and human machine interface. Despite recent innovations, real-time multiple objects tracking has remained one of the most challenging problems in the wide range of computer vision applications. In this paper we propose a tracking-by-detection method that builds up a weighted scoring mechanism and associate the detected vehicles to the trackers by maximizing the score function for the vehicle tracking and counting purpose. With the help of our PVA-lite deep learning model as the vehicle detector, the counting accuracy of the proposed method can reach more than 95% with the system performance of 5fps @320x240 on Nvidia Jetson TX1 including vehicle detection, tracking and counting.

II. PROPOSED METHOD

2.1 Overview

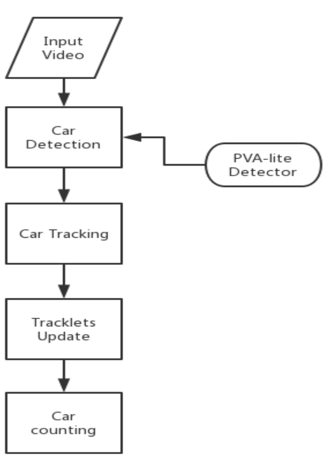

Figure 1. *Flowchart of the proposed method*

Figure 1 shows the proposed system flowchart. The proposed tracking-by-detection algorithms will be described in Section 2.3.

2.2 Foreground extraction

Foreground extraction is a process that can differentiate between foreground and background by comparing input frame and the background model. The foreground is regarded as containing people. Background subtraction is performed to retrieve the difference between input and background model in every pixel. Each pixel is categorized as foreground or background by binarization.

2.2.1 False Positive Elimination

Here, we define a ratio for every bounding box

$$\boldsymbol{R_b} = \frac{\sum_{x=0}^{h_x}\sum_{y=0}^{h_y} \boldsymbol{Binarization(x,y)}}{\boldsymbol{h_x \cdot h_y}} \qquad (1)$$

where h_x and hy are the width and height of a bounding box respectively. We define a threshold $T_b(=0.15)$. If R_b is smaller than T_b, the bounding box will be regarded as a false positive.

2.3 Tracking Algorithm

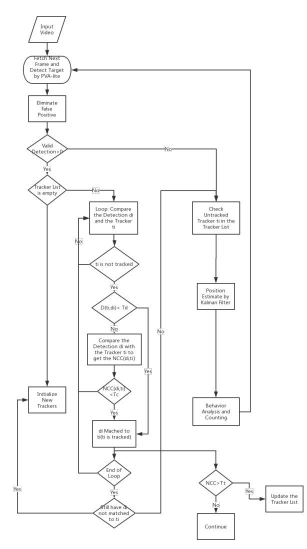

Figure 2. Flowchart of Tracking Algorithm

2.3.1 Normalized Cross Correlation (NCC)

NCC is a value to compare two distributions to see how much they are correlated. Image plane can be viewed as a two dimensional distribution. Therefore, by substituting the images for the distributions, NCC can be seen as the similarity of two images. We introduce the NCC equation as shown in Eq. (2).

$$\mathbf{NCC} = \frac{\sum_{y=0}^{h_y-1}\sum_{x=0}^{h_x-1}[T(x,y)-\mu_T][C(x,y)-\mu_C]}{\sqrt{\sum_{y=0}^{h_y-1}\sum_{x=0}^{h_x-1}[T(x,y)-\mu_T]}\sqrt{\sum_{y=0}^{h_y-1}\sum_{x=0}^{h_x-1}[C(x,y)-\mu_C]}} \qquad (2)$$

2.3.2 Data association

In Eq. (3), we define a weighted score between the detection and the trackers.

$$s_j^i = W_{ncc} \cdot NCC(d(i), t(j)) + W_{euc} \cdot EUC(d(i), t(j)) + W_{area} \cdot AREA(d(i), t(j)) \qquad (3)$$

s_j^i means the score of detection *d(i)* and the tracker *t(j)*. W_{ncc}, W_{euc}, and W_{area} are weights for NCC, Euclidean distance (in pixels), area overlap ratio between the detection and tracker respectively. Only those pairs whose s_j^i is bigger than a preset threshold T_c will be connected. As illustrated in Figure 3, the green dots and the red dots are the successfully matching pairs. The purple dot doesn't have any match in trackers so it will initialize a new tracker. The red dot leaves unconnected, so Kalman filter will be implemented on it.

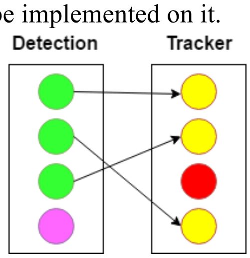

Figure 3: Data association in the proposed algorithm

2.3.3 Tracking process

After fetching the video sequences, the detector will use the sliding window strategy to detect the car. Then we use data association method to match between the detection results and the current trackers. A new tracker will be initiated by *d(i)* if there is no match between *d(i)* and the trackers.

Our tracking mechanism consists of three essential parts. First, initialization, which is given by the detector. Second, data association, which is the process of matching detection results to the current trackers. Third, prediction, which will predict the position of the lost tracker by Kalman filter. In the data association step, each pair of detection and tracker will be given a score by weighting their appearance, distance, size, and motion. The details of tracking process will be showed in Figure 2.

III. EXPERIMENTAL RESULTS

This section shows the experimental results of our proposed system. We test the proposed algorithm on some video clips. In Table 1, it shows the average counting accuracy can reach more than 95%.

Table 1: Counting accuracy

#	Name	Scene	GroundTruth (Target Number)	Counting Result	Accuracy Rate
1	High3	Highway	63	63	98.4%
2	High4	Highway	52	52	100%
3	Rain	Rainy-Highway	94	87	93%

Table 2: Nvidia Jetson TX1 specification

Processor	64-bit ARM Cortex-A57 Quad-core
Memory	4GB LPDDR4
Video In	File input
Operation System	Linux Ubuntu 14.04

IV. CONCLUSION

We have proposed a tracking-by-detection method for single surveillance camera, which includes data association by weighting the similarity score of the detection and the trackers. And for the unconnected trackers, we proposed to use Kalman filter to predict the position. The overall counting accuracy can reach more than 95% and the system performance can achieve 5fps @320x240 on Nvidia Jetson TX1.

REFERENCES

[1] E. Rosten, R. Porter, and T. Drummond, "Fusing points and lines for high performance tracking," *IEEE International Conference on Computer Vision (ICCV)*, 2005.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, pp 91–110, 2004.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *European Conference on Computer Vision (ECCV)*, 2006.

[4] M. Calonder, V. Lepetit, C. Strecha, P. Fua, "BRIEF: Binary robust independent elementary features," *European Conference on Computer Vision (ECCV)*, 2010.
E. Rublee, V. Rabaud, K., Konolige and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[5] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," *IEEE International Conference on Computer Vision*, 2011.

[6] A. Alexandre, R. Ortiz, and P. Vandergheynst, "Freak: Fast retina keypoint," *IEEE Conference on Computer Vision and Pattern Recognition (CCVPR)*, 2012.

[7] D. Merad, K.-E. Aziz, R. Iguernaissi, B. Fertil, and P. Drap, "Tracking multiple persons under partial and global occlusions: Application to customers' behavior analysis," *Pattern Recognition Letters*, vol. 81, pp. 11-20, 2016

[8] F. Bousetouane, L. Dib and H. Snoussi, "Improved mean shift integrating texture and color features for robust real time object tracking," *The Visual Computer*, vol. 29, pp. 155-170, 2013

[9] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011

[10] B. Yang and R. Nevatia, "Online learned discriminative partbased appearance models for multi-human tracking," *European Conference on Computer Vision (ECCV), 2012*