An End-to-End Traffic Vision and Counting System Using Computer Vision and Machine Learning: The Challenges in Real-Time Processing

Haiyan Wang, Mehran Mazari, Mohammad Pourhomayoun
Computer Science Department
California State University Los Angeles
Los Angeles, USA
Email: mpourho@calstatela.edu

Hunter Owens
Data Science Federation
City of Los Angeles
Los Angeles, USA
Email: hunter.owens@lacity.org

Janna Smith
Department of Transportation
City of Los Angeles
Los Angeles, USA
Email: janna.smith@lacity.org

William Chernicoff
Toyota Mobility Foundation
Washington DC, USA
Email: william.chernicoff@toyota.com

Abstract— **The goal of this research is to design and develop an end-to-end system based on computer vision and machine learning to monitor, count, and manage traffic. The end goal of this study is to make our urban transportation safer for our people, especially for pedestrians and bicyclists, who are the most vulnerable components of traffic collisions. Several methods have been proposed for traffic vision, particularly for pedestrian recognition. However, when we want to implement it in real-time in the scale of a large city like Los Angles, and on live video streams captured by regular traffic cameras, we have to deal with many challenges. This paper introduces the main challenges in traffic vision in practice, and proposes an effective end-to-end system for traffic vision, detection, tracking, and counting to address the challenges.**

Keywords - Computer Vision; Machine Learning; Kalman Filter; Object Detection.

I. INTRODUCTION AND MOTIVATION

More than 50% of the world's population now live in urban areas. By 2050, 66% of the world's population is projected to be urban [1][2]. As urban populations rise, it is essential for city planners and designers to focus more on designing smart cities and addressing the main challenges such as traffic issues, and the impacts of increased vehicle use.

According to the U.S. Department of Transportation (USDOT), the number of traffic fatalities has increased by nearly 6% in 2016 [3]. The city of Los Angeles has one of the highest rates of traffic death among large U.S. cities. Every year, more than 200 people die in traffic accidents only in the city of Los Angeles. The most vulnerable components of the traffic collisions are pedestrians and bicyclists (accounted for almost half of the fatalities). Thus, it is essential to develop intelligent transportation systems, and human-centered traffic approaches to protect our pedestrians and bicyclists and ensure that they can travel safely, efficiently, and comfortably to their destinations.

The goal of this study is to design and develop an end-to-end system based on computer vision and machine learning to monitor, track, count, and manage traffic, particularly to monitor and count pedestrians and bicyclists in real-time.

Traffic monitoring has become a very popular and important field of research and study in the past couple of years. Several methods have been proposed for traffic vision, particularly for pedestrian detection [4][5]. However, when we want to do it in practice, in real-time on video streams captured by regular traffic cameras, and when we want to implement it in the scale of a large city, it will be very different from lab settings and we have to deal with many challenges.

This paper particularly introduces the main practical challenges in traffic vision, and proposes an effective end-to-end system for traffic vision, detection, tracking, and counting, and addresses the main challenges in this field.

II. PRACTICAL CHALLENGES IN TRAFFIC VISION

Here are some of the main challenges and difficulties that a traffic vision system may face:

- Poor quality of videos because of camera low resolution, light conditions, dirty or unadjusted lens, or weather conditions.
- Dealing with stretched, convex, or squeezed videos collected by traffic cameras and wide-angle lenses.
- Undesired angle, location, and direction of the camera.
- Camera vibration and shake because of wind or the cars passing by.
- Light distortion at night by passing cars.
- Inconsistent light change during the day time and shadow effect.
- Moving or stationary objects that may block the target view.

 ISBN: 978-1-61208-638-5

Figure 1 illustrates some examples of difficult situations. In Figure 1-(a), the pedestrian is partially visible. The view is blocked by the wall. In Figure 1-(b) the light condition is poor and inconsistent, it is even difficult to detect the pedestrian with human eye. The video is stretched and convex (as we see, the curb line in the left is completely converted to a curve). In Figure 1-(c), the pedestrian is hardly visible because of extreme shadow and inconsistent light.

(a)

(b)

(c)

Figure 1. Examples of difficult situations: (a) Undesired location and direction of camera, and also camera vibration; (b) Light condition is poor and inconsistent, also the video is stretched and convex; (c) Extreme shadow, light change during the day time, inconsistent lighting.

III. System Architecture and Method

We have developed an end-to-end system including a series of image/video processing, computer vision algorithms, Machine Learning, and optimal state estimator algorithms that receive video streams in real-time, and monitor, recognize, track, and count pedestrians and cyclists in the video. The next sections describe the 3 main parts of the system: 1) raw video processing, 2) feature engineering and Machine Learning for object detection, and 3) trajectory prediction for traffic tracking and counting.

A. Video Processing

Figure 2 shows the system architecture. The first step in an end-to-end traffic vision system is raw video preprocessing, which includes a series of standard algorithms for quality enhancement, and brightness and contrast adjustment. In the case of wide-angle lenses that may make the image convex, we can also use correction algorithms to convert the video back to more natural view.

The next step in our system is background estimation and subtraction (we can also call it foreground detection or moving object detection). In this concept, any moving object is considered as foreground, and any stationary object in a period of time (i.e., an object with fixed location in a number of sequential frames) is considered as background. We have tried several effective algorithms for background estimation/subtraction including frame differencing, mean filter, running Gaussian average, and mixture of Gaussian modeling (MOG) [6][7]. It turned out that mixture of Gaussian modeling (MOG), and also mean filtering achieved the best results for background subtraction. Figure 3 shows the results of background subtraction (i.e., moving object detection) based on mean filtering.

It is important to notice that the background continuously changes during the day-time because the sunlight direction and intensity changes. Figure 4 shows the estimated background of a video captured by a traffic camera at 7:06AM and another time at 7:19AM. As we see, the background has significantly changed in only 13 minutes. Thus, we need to continuously estimate and update the

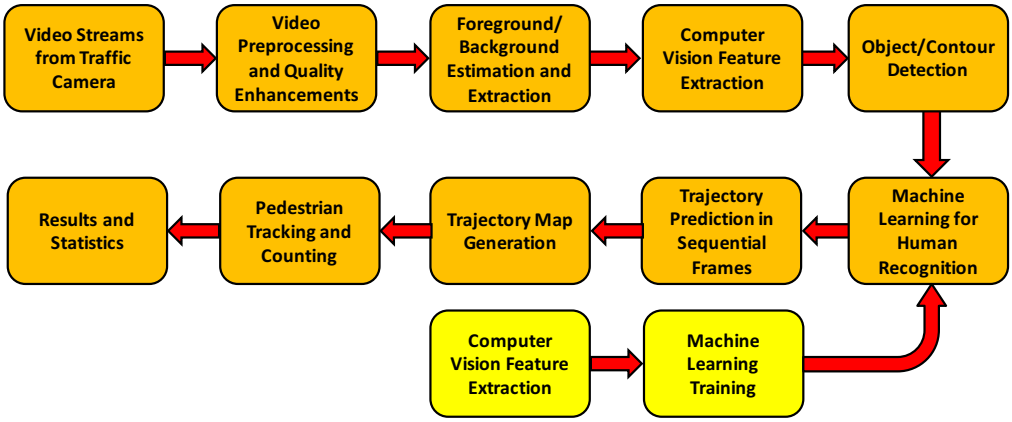


Figure 2. End-to-end system architecture.

background to always have the best background subtraction performance.

We have to notice that background removal not only improves the performance and accuracy of object recognition algorithm (i.e., the next step, which is machine learning algorithm), but also significantly reduces the computational load of the object recognition algorithm by reducing the size of the area of interest. This will be even more important when we want to use computationally expensive machine learning algorithms such as Convolutional Neural Networks (ConvNet also known as CNN) [13].

(a)

(b)

(c)

Figure 3. Background subtraction: (a) Original video frame, (b) Estimated background, (c) Moving objects after background subtraction.

Figure 4. Background change in only 13 minutes.

B. *Machine Learning for Object Detection*

The next step is to extract and select the best set of computer vision features that can be used in machine learning algorithms for object detection. Depending on the type of machine learning algorithm, this step may include feature extraction, feature selection, and/or dimensionality reduction. We have tried many different types of features and machine learning algorithms for object recognition.

One of the most effective and popular features are Histogram of Oriented Gradient (HOG) features [8] that along with SVM classifier can form an effective method for pedestrian recognition [8]. HOG is a feature descriptor that counts occurrences of gradient orientation in localized portions of an image [8]. It has been proven to be one of the most effective hand-made features that can be used for object recognition.

We have also tried deep learning methods, particularly the Convolutional Neural Networks (ConvNet) including R-CNN (Region-based Convolutional Network), and YOLO (You Only Look Once) algorithms [9][10][11]. A big advantage of ConvNet methods compared to other classic machine learning algorithms is that there is no need to generate and use hand-made features for ConvNet. The algorithm automatically learns to generate the best set of convolutional features that can best represent the image. However, ConvNet is computationally expensive and sometimes very difficult to run in real-time on high-frame-rate videos. In addition, when the training dataset is not large enough, it is usually hard to train a deep neural network. In this case, *Transfer Learning* methods that take advantage of a pre-trained neural network model on another dataset can be very helpful to ease and expedite the training stage.

Figure 5-(a) shows our pedestrian detection results using HOG features and SVM classifier. Figure 5-(b) shows our results using YOLO algorithm.

(a)

(b)

Figure 5. Pedestrian detection using machine learning algorithms. (a) using HOG features and SVM classifier, (b) using YOLO.

C. *Trajectory Prediction for Traffic Tracking and Counting*

After detecting a target object (e.g., a pedestrian or bicyclist) in several sequential frames, we use *Optimal State Estimator* to estimate the *Trajectory* of each target object. Since several objects may exist in each frame at a time (e.g., several pedestrians walking together in same direction or different directions), it is essential to estimate the trajectory of each object individually.

To this end, we use Kalman Filter [12] as an optimal state estimator to predict the next location of the object and estimate the trajectory of the object over time. This allows us to track each object individually during the video. For example, suppose that we want to track a pedestrian. We use Kalman filter to predict the next location of the pedestrian in next frame based on its previous locations and walking pace. Then, after receiving the next frame, we compare our prediction with the actual pedestrian detected in next frame. This comparison tells us if this pedestrian was the same person in previous frame, or it is a new one. If the predicted location and actual location match, we consider this pedestrian as previous one, and continue completing the trajectory of this pedestrian (see Figure 6). Using this approach, we can build a trajectory map including individual trajectories for all pedestrians in the video, and then track each pedestrian from the first frame he enters until the last frame when he moves out.

Every time we detect a pedestrian whose location does not match to any of the previously predicted locations (i.e., it does not locate on any of the existing estimated trajectories), we consider that person as a new pedestrian and consequently, increment the pedestrian counter. This will allow us to track and count each pedestrian everywhere in the video, and avoid double counting them in sequential frames.

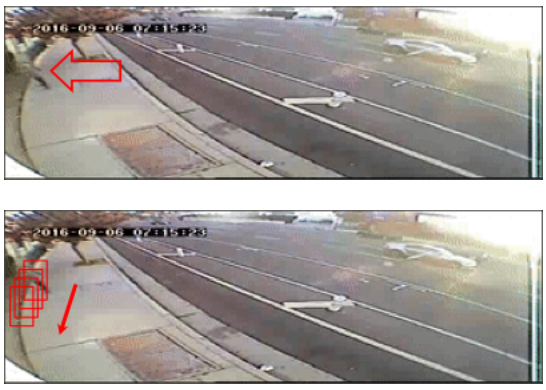
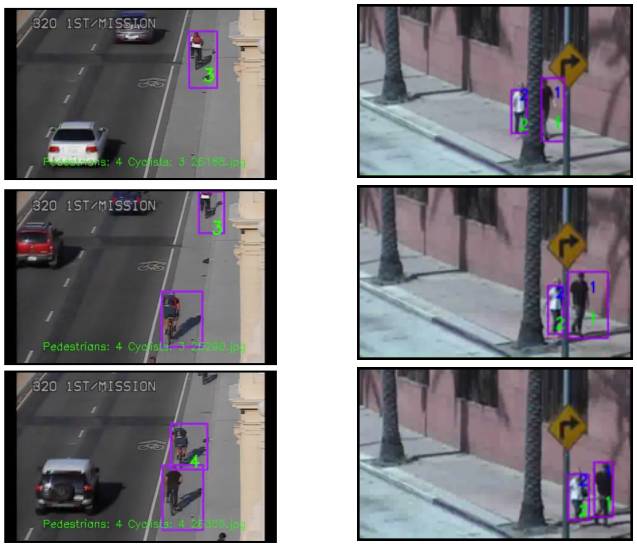


Figure 6. Location prediction and Trajectory estimation.

IV. CONCLUSION

The city of Los Angeles has one of the highest rates of traffic death among large U.S. cities. Fortunately, the city has launched the *Vision Zero* initiative as a strategy and commitment to reduce traffic fatalities. The main goal of LA's Vision Zero is to eliminate all traffic fatalities by 2025. Since the most vulnerable components of traffic accidents are pedestrians and bicyclists, it is essential to develop intelligent transportation systems, and human-centered traffic approaches to protect pedestrians and bicyclists. This paper introduced an effective end-to-end system based on computer vision and machine learning to detect, monitor, track, and count pedestrians and bicyclists in real-time. This approach particularly enables us to recognize and monitor busy intersections that are prone to traffic accidents, and allows us to control and manage traffic in those intersections to protect our pedestrians and bicyclists.

The California State University Los Angeles in partnership with the Los Angeles Department of Transportation (LADOT), the City of Los Angeles' Data Science Federation, and Toyota Mobility Foundation has developed this effective and scalable system to detect, monitor, track, and count pedestrians and bicyclists in real-time. This system is potentially scalable to the 56,000 miles of streets in Los Angeles. Despite of many practical challenges (mentioned in Section II), the developed system works very well with the existing regular traffic cameras and therefore, there is no need to install any special or new cameras for this purpose. This system will help to increase safety and traffic flow through better traffic management and planning. This would be transferrable to other cities and municipalities as well. Figure 7 shows some of the results for pedestrian and bicyclist detection, tracking, and counting on real video streams captured by traffic cameras in Los Angeles.



(a) (b)

Figure 7. System results on real-time traffic video streams: (a) Bicyclist tracking and counting, (b) Pedestrian tracking and counting.

ACKNOWLEDGMENT

The authors would like to appreciate Toyota Mobility Foundation for supporting this research. The authors would like to appreciate LADOT, City of LA, and ITA Data Science Federation for valuable help and support.

REFERENCES

[1] World Urbanization Prospects, UN-Department of Economic and Social Affairs.

[2] Unicef, Urban World, www.unicef.org/sowc2012/urbanmap.

[3] USDOT, https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," in IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012.

[5] Benenson R., et al. “Ten Years of Pedestrian Detection, What Have We Learned?” ECCV, Springer, 2015.

[6] M. Piccardi, “Background subtraction techniques: a review”, IEEE Int. Conf. on Systems, Man and Cybernetics, 2004.

[7] T. Bouwmans, F. El Baf, and B. Vachon, "Background Modeling using Mixture of Gaussians for Foreground Detection – A Survey". Recent Patents on Computer Science, 2008.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[9] R. Girshick, "Fast R-CNN," IEEE International Conference on Computer Vision (ICCV), 2015.

[10] S. Ren, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," IEEE Conf. on Computer Vision and Pattern Recognition, 2016.

[12] P. Zarchan and H. Musoff, “Fundamentals of Kalman Filtering: A Practical Approach”, ISBN 978-1-56347-455-2, 2000.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, NIPS 2012.

 ISBN: 978-1-61208-638-5