Decoding the Spatiotemporal Scene of a Road-Traffic Intersection for Real Time Event Detection

Justin A. Eichel, Akshaya Mishra, Nicholas Miller, Nicholas Jankovic, and Kurtis McBride
All authors have contributed equally
Miovision Technologies, 101-148 Manitou Dr, Kitchener, ON N2C 1L3, Canada

Abstract

As large-scale, city-wide sensor networks integrate with emergency response services, regional maintenance crews, and traffic control systems, the volume of sensory data necessitates the use of event recognition for real-time information routing and control. The proposed work describes a unified computer vision framework that employs spatiotemporal reasoning and scene modeling to detect stationary objects and non-stationary events at traffic intersections. Video imaging vehicle detection systems (VIVDS) are deployed at 22 intersections to collect traffic video data. The success of real-time roadside event detection and classification is highly dependent on learning the most representative event features, which have been extracted from approximately 200 hours of annotated video using an offline distributed cloud-based training infrastructure. The median accuracy of the detection system is 92%.

1. Background

Being able to understand a scene to extract meaningful events is of great interest to researchers in such fields as activity detection, surveillance, traffic parameter estimation and navigation [1]. Many scene understanding techniques [2] have described static scenes for applications in content based image and video retrieval.

Video imaging vehicle detection systems (VIVDS) are now common in the traffic industry [3], where vehicle detection typically employs background subtraction and blob tracking. Simple implementations have drawbacks that may include false vehicle detections due to lighting changes and "ghosting" in the background subtraction [3, 4]. Furthermore, most VIVDS have strict constraints on scene perspective, necessitating the installation of multiple cameras for one intersection. Multiple cameras increase capital and maintenance cost, while making deployments more prone to error [3]. The authors have utilized scene understanding to improve the reliability and performance of real-time traffic event detection aiming to meet industry expectations. For example, the Florida Department of Transportation states that a VIVDS must measure vehicle presence, volume, speed and occupancy with > 95% accuracy [5]. A VIVDS must also detect vehicles that are up to 300 ft. away. Other agencies, such as the California Department of Transportation, also require VIVDS to detect cyclists and pedestrians.

This application focuses on detecting and counting vehicles, cyclists, and pedestrians. The objective is to create a VIVDS that can identify various traffic events in real-time. While vehicle detection is almost trivial under favorable conditions, additional scene information is needed to track and classify vehicles under changing real-world conditions. Accurate detection requires estimates for expected vehicle location and size, direction of traffic flow, probabilities of class occurrence, and static occlusions within the scene.

System Description: This traffic event detection system consists of two major components: a VIVDS and a supervised learning and training infrastructure. Each intersection employs one VIVDS, which is composed of a high-resolution fisheye camera that feeds video to a ruggedized multi-core processor. The processor records video for later (offline) analysis and processes it in real-time to generate vehicle detections that are relayed to the roadside traffic control system. The learning and training system currently uses more than 100,000 annotated objects from the recorded training videos to compile a set of over 10,000 features each and extracts around 100 important features that best classify objects into classes and subclasses such as cars, trucks, road, and cyclists.

2. Scene Event Understanding

The VIVDS employs several scene parameter event estimation methods, including image rectification, online medoid-based background modeling, dynamic foreground modeling, dynamic traffic lane estimation, online stationary foreground occlusion handling, in-painting, and utilization of prior object statistics.

Rectification: Non-linear warping and scaling makes training an object classifier on a circular fisheye image difficult. Intra-class object scale and pose variations are mitigated by homogenizing the viewpoint of each region of interest (ROI) so that all vehicles are scale invariant along their trajectories and travel horizontally, from left to right. See Figure 1. Methods are also available to take any image point and map them to real-world GIS coordinates.

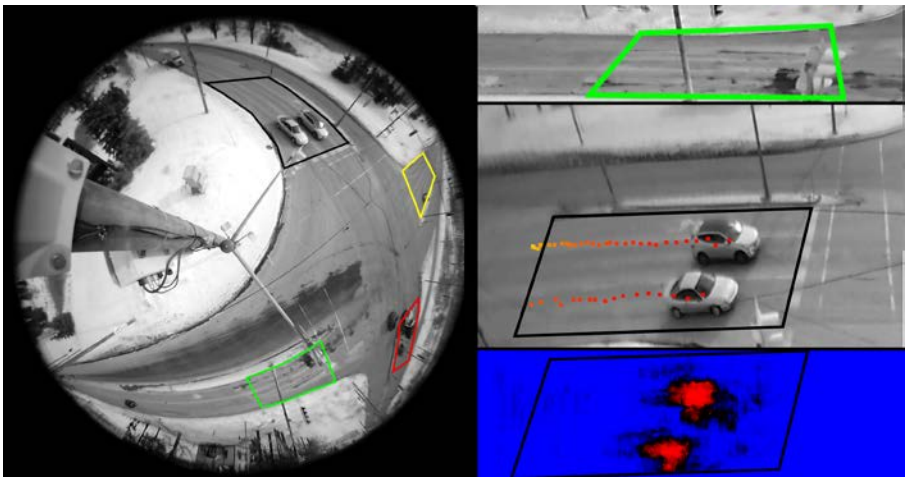

Figure 1: A spherical image (left) captured from the VIVDS is rectified using a non-linear transformation. The rectification (right) ensures incoming traffic has a homogenized flow from left to right and vehicles have constant height for better event recognition. The event detector (middle right) tracks two vehicles (one travels through the intersection and the other prepares to turn right). It employs a sliding window classification map (bottom right, blue = road, red = car).

Prior event statistics: Object and event statistics for a scene are collected and analyzed to produce a statistical machine learning classifier. The a priori probability of an object of specified height, width, and class, traveling in a specific lane is calculated from sample statistics to improve event-lane assignment accuracy.

Dynamic background and foreground model: Since it is not feasible to search all positions and scales for multi-class and multi-scale spatiotemporal events in real-time, a medoid-based background model is employed to reduce the potential event search space. Region based statistics allow the background model to adapt to rapid changes in weather and lighting by quickly absorbing environmental changes that affect a particular neighborhood.

Dynamic traffic lane boundaries: Vehicles can change their preferred lane alignment at an intersection due to adverse weather conditions or roadway obstructions. Dynamic lane boundary estimation via maximum likelihood from data clustering of spatiotemporal detector hits improves accuracy in such cases. The density of detector responses over time provides up-to-date information on likely image positions of a traffic lane.

3. Discussion

Scene understanding improves the robustness of vehicle detections in cases where lane geometry and environmental conditions change over time. Machine learning continually improves detection performance resulting in an accurate and affordable product. Further, detections can be erroneously reported in the vicinity of static foregrounds, such as a lamp posts or sign mast arms. Some lower-level image processing algorithms are unable to overcome such obstructions. Rather than reconstituting a detected object from multiple segments, in-painting [6] allows the occluded parts of an object to be estimated using that object's predicted trajectory; thus allowing uninhibited image processing behind occlusions.

The advantages of scene understanding are achievable at a real-time 15 fps performance. On average the mediod based foreground estimation and the sliding window classifier require 8 and 29 msec, respectively, using a 3.2 GHz AMD CPU with 4GB of RAM.

References

[1] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele. Monocular Visual Scene Understanding: Understanding Multi-Object Traffic Scenes, TPAMI, IEEE, 35(4):882-897, 2013.

[2] J. Vogel and B. Schiele. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. Int. Journal of Computer Vision, 72(2), 2007.

[3] J. Medina, R. Benekohal, and M. Chitturi. Evaluation of Video Detection Systems: Effects of configuration changes in the performance of video detection systems. Illinois Center for Transportation, Issue 8, Part 24 of Civil Engineering Studies, 2008.

[4] O. Barnich, and M. Van Droogenbroeck. ViBe: A Universal Background Subtraction Algorithm for Video Sequences. TIP, IEEE, 20(6):1709-1724, 2011.

[5] Florida Department of Transportation. Standards Specifications for Road and Bridge Construction, Vehicle Detection System, 832-839, 2014.

[6] K.A. Patwardhan, G. Sapiro, and M. Bertalmio. Video Inpainting Under Constrained Camera Motion, TIP, IEEE 16(2):545-553, 2007.