AUTOMATED TRAFFIC SURVEILLANCE FROM AN AERIAL CAMERA ARRAY

FINAL REPORT

SOUTHEASTERN TRANSPORTATION CENTER

WAYNE SARASUA, XI ZHAO, DOUG DAWSON,

AND STANLEY BIRCHFIELD

JULY 2016

US DEPARTMENT OF TRANSPORTATION GRANT DTRT13-G-UTC34

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Automated Traffic Surveillance from an Aerial Camera Array		5. Report Date July 2016
		6. Source Organization Code Budget
7. Author(s) Sarasua, Wayne; Zhao, Xi; Dawson, Douglas; and Birchfield, Stanley		8. Source Organization Report No. STC-2015-##-XX
9. Performing Organization Name and Address Southeastern Transportation Center UT Center for Transportation Research 309 Conference Center Building Knoxville TN 37996-4133		10. Work Unit No. (TRAIS) 11. Contract or Grant No. DTRT13-G-UTC34
12. Sponsoring Agency Name and Address US Department of Transportation Office of the Secretary of Transportation–Research 1200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered Final Report: June 2014-July 2016 14. Sponsoring Agency Code USDOT/OST-R/STC

15. Supplementary Notes:

16. Abstract

The research focuses on a novel computer vision-based traffic surveillance system capable of processing aerial imagery to track vehicles and their movements. The system uses a preprocessed 1-Hertz image sequence with a coverage of 25 square miles from an aerial camera array mounted on an airplane. The unique characteristics of the input data make this work challenging. Several heuristic and machine learning approaches are proposed and evaluated to detect and track vehicles with the purpose of collecting different roadway traffic parameters including speed, density, and volume. The research has potential to be useful for "big data" monitoring of traffic parameters over an entire 25 mi^2 area with a single sensor. The researchers tested and evaluated a number of different computer vision approaches to solve the detection and tracking problem and discovered that deep learning combined with SURF-based approaches provided the best results. The system achieved over 93% accuracy in traffic density estimates and over 90% accuracy in speed estimates on 50 seconds of data when compared with manually collected ground truth. It has 100% accuracy when level of service (LOS) was calculated for several uninterrupted flow facilities tested. These evaluations were conducted for facilities of different levels of congestion. In addition to traffic data collection, the sensing system is capable of identifying traffic incidents by monitoring abrupt increases in traffic density. With further research, improved preprocessing, and a higher frame rate, the accuracy of tracking vehicles can be improved which will allow for other potential applications such as identification of erratic drivers and origin-destination studies.

17. Key Words Traffic Surveillance, Aerial Camera Arrays, Vehicle Tracking, Vehicle Detection, Computer Vision, SURF, Deep Learning	18. Distribution Statement Unrestricted; Document is available to the public through the National Technical Information Service; Springfield, VT.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages #	22. Price …

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Traditional traffic sensors (e.g., inductive loop detectors, microwave radar, infrared devices, piezos, and road tube sensors) are used to monitor traffic and collect data at fixed points throughout a traffic network. These data include traffic volume, time-mean speed (average speed at a point), vehicle classification, and occupancy. The capability to track vehicles through a network with these sensors is not possible. However, such a capability has broad traffic applications including identification of reckless or impaired drivers and traffic data collection. With the advent of aerial digital camera arrays, it is now possible to record high-resolution video for a wide field of view from aircraft overhead.

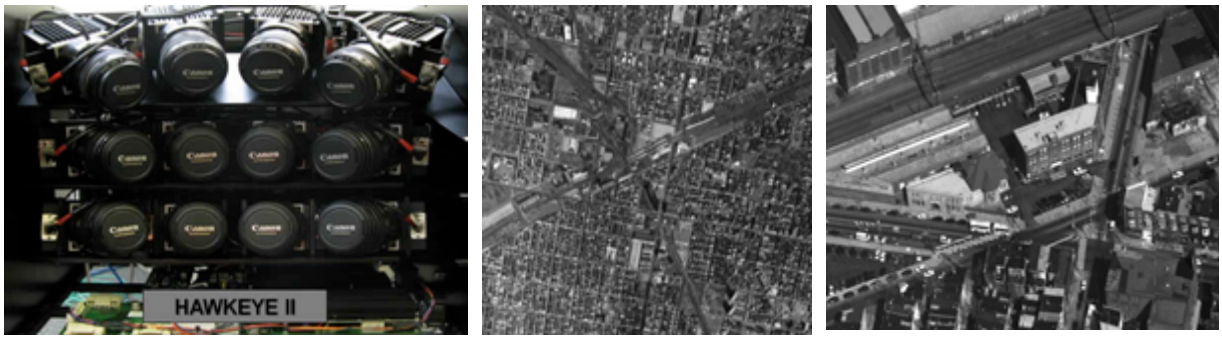
The objective of this research was to develop an automated traffic surveillance system capable of processing aerial camera array imagery to extract valid and useful traffic data for diverse applications including traffic data monitoring and traffic safety. To achieve this objective, we employ digital image processing and computer vision techniques to explore feasble approaches and improve their performance. The system uses a preprocessed 1-Hertz image sequence with a coverage of 25 square miles from an aerial camera array mounted on an airplane. The unique characteristics of the input data make this work challenging. Several heuristic and machine learning approaches are proposed and evaluated to detect and track vehicles with the purpose of collecting different roadway traffic parameters including speed, density, and volume. The research has potential to be useful for "big data" monitoring of traffic parameters over an entire 25 square mile area with a single sensor.

The researchers tested and evaluated a number of different computer vision approaches to solve the detection and tracking problem and discovered that deep learning combined with SURF-based approaches provided the best results. The system achieved over 93% accuracy in traffic density estimates and over 90% accuracy in speed estimates on 50 seconds of data when compared with manually collected ground truth. It has 100% accuracy when level of service (LOS) was calculated for several uninterrupted flow facilities tested. These evaluations were conducted for facilities of different levels of congestion. In addition to traffic data collection, the sensing system is capable of identifying traffic incidents by monitoring abrupt increases in traffic density. With further research, improved preprocessing, and a higher frame rate, the accuracy of tracking vehicles can be improved significantly which will eventually allow the envisioned system to be able to accurately map the location of vehicles throughout a camera array image sequence. By mapping the locations of vehicles in a spatio-temporal manner, additional traffic parameters can be extracted microscopically. These may include turning movements at intersections, traffic signal timings, identification of erratic and drunk drivers; and origin-destination data. Further, such a system could facilitate traffic management through traffic data mining. Traffic data mining methods provide a means to generate dependable patterns from large traffic data sets with less complexity than current approaches. When traffic predictions are accurate, more efficient management of the traffic network is possible. A digital real-time "traffic map" created by the envisioned system will provide a robust data set where data mining methods could be applied to enhance traffic management and provide data for a variety of traffic studies.

DESCRIPTION OF PROBLEM

Traditional data collection sensors (e.g., inductive loop detectors, microwave radar, infrared devices, piezos, and road tube sensors) are used to monitor traffic and collect data at fixed points throughout a traffic network. These data include traffic volume, time-mean speed, vehicle classification, and occupancy. Such sensors, however, are not able to track vehicles through a network, even though such a capability has broad traffic applications by providing microscopic parameters of individual vehicles. Vehicle tracking can provide stopped delay, running speeds, acceleration and deceleration, and other driver behavior characteristics. Many systems with cameras mounted on the roadside were developed to collected traffic image data for various applications including collision detection (Saunier and Sayed 2007) or driving behavior (Tsai et al. 2011). With the advent of aerial digital camera arrays, it is now possible to record high-resolution video for a wide field of view from aircraft overhead. Fig. 1 (a) shows such a camera array that is configurable for a variety of aircraft. The area of coverage of this system is altitude dependent, but the video we processed that was captured by this particular device covers an area of approximately 8.05 km by 8.05 km (5 miles by 5 miles), as shown in Fig. 1 (b) and (c).

Automated processing of video from a camera array to extract traffic data has yet to be accomplished because of the unique challenges involved. These challenges include video stabilization, image registration and rectification, object recognition, and low-frame-rate tracking. The objective of this research was to develop an automated traffic surveillance system capable of processing aerial camera array imagery to extract valid and useful traffic data for diverse applications including traffic data monitoring and traffic safety. To achieve this objective, we employ digital image processing and computer vision techniques to explore feasble approaches and improve their performance.



(a) (b) (c)

Fig. 1. (a) Aerial camera array; (b) High-resolution image; (c) Zoomed view (www.pss-1.com)

Previous Work on Digital Image Processing of Aerial Video of Traffic

Many algorithms have been developed to detect vehicles automatically in aerial images, but most of the testing was done using images from a single camera or captured at low altitudes with a small field of view. For example, researchers at the University of Arizona used a

computer vision-based approach to collect traffic parameters from a single low-resolution camera (720 × 480) mounted to a helicopter flying at an altitude of under 305 m (1000 feet), which provided a field of view of less than 244 m (800 feet) (Angel et al. 2002). Their continued research led to the development of a software called "Tracking and Registration of Airborne Video Image Sequences" (TRAVIS) which can extract vehicle positions from airborne imagery to assist the analysis of microscopic traffic behavior. The input of TRAVIS is a sequence of images from captured airborne video. TRAVIS registers the image sequence to an initial common reference frame, detects the vehicles in the images and tracks the vehicles through the image sequence using a blob tracking algorithm. The output of TRAVIS is a sequence of pixel coordinates of vehicles as they are tracked through the image sequence (Hickman and Mirchandani 2006). Follow-up research was conducted by Du and Hickman (2012) to improve vehicle detection and reduce the probability of false detection and computation time by masking areas outside roadways. They also improved the tracking algorithms to better handle vehicles with little contrast relative to the pavement of roadways. The dataset used for this recent work was collected by a single camera from a helicopter that provided a 0.4-m (1.3-ft) pixel size.

Most approaches in previous work on vehicle detection and tracking can be categorized into one of three classes: feature detectors, background subtraction/modeling and machine learning. Many different feature detectors, such as KLT and SIFT, are widely used to either track vehicles or model the appearance of vehicles for detection and tracking purposes. Moon et al. (2002) designed a vehicle detector by combining four elongated edge operators. The performance was affected by camera angles and illuminations. Kim and Malik (2003) presented a model-based 3D vehicle detection and description algorithm based on line features, and the algorithm outperformed Zhao and Nevatia's algorithm (2003). Hinz (2005) modeled cars on a local level with a 3D wireframe representation and on a global scale by the grouping of cars within queues; this approach did not rely on external information and was not limited to constrained environments. Palaniappan et al. (2010) presented an interactive tracking system based on feature detection using appearance modeling and motion prediction. Cao et al. (2012) proposed a framework for vehicle detection and tracking robust to partial occlusion based on KLT features. Pelapur et al. (2012) presented a feature tracking system which used an adaptive set of feature descriptors with posterior fusion modeling.

Background subtraction and modeling are used frequently either with a stabilized camera or with a system for which an accurate image registration is possible. Reinartz et al. (2006) used background subtraction to find vehicles and image patch correlation to match the vehicles between frames. Their approach had issues with mistakenly detecting pedestrians and grouping vehicles that were too close together. Xiao et al. (2010) proposed a probabilistic relation graph to combine a vehicle behavior model with a road network for vehicle detection and tracking in wide area video. Shi et al. (2013) proposed a maximum consistency context model to assist background subtraction based multiple object tracking by leveraging the discriminative power and robustness in the scenario. Prokaj and Medioni (2014) presented a multiple object tracking approach using two trackers in parallel: one based on detection by background subtraction and the other one using a template based regression tracker. Saleemi and Shah (2014) presented a framework capable of tracking thousands of vehicles in low frame rate aerial videos using background modeling. Chen and

Medioni (2015) developed two methods of adapting the background model for more accurate background subtraction even when there is parallax (for example with buildings). The first method used a dense 3D model of the landscape and the second method made use of the epipolar flow constraint.

Machine learning is widely used in vehicle detection along with either feature detectors or background subtraction/modeling. Zhao and Nevatia (2003) introduced a passenger car detection system by modeling passenger cars as 3D objects using a Bayesian network. Nguyen et al. (2007) developed an automatic car detection framework using AdaBoost which was trained with three types of features. Tuermer et al. (2010) used a preprocessing algorithm to limit the search space and developed a reliable detector using Real AdaBoost with HoG features. Cheng et al. (2012) introduced a pixelwise classification approach in which a dynamic Bayesian network (DBN) was constructed for classification.

Most vehicle tracking approaches are based on vehicle detections, which utilize the visual information to initialize the tracker or support the tracking process to match correspondences between adjacent frames. However, in some approaches, the tracking process and detection process are mutually dependent. Kalal et al. (2012) proposed the TLD framework which combined tracking, learning and detection; the tracker generates training data for improving the detector and the detector initializes and re-initializes the tracker simultaneously.

None of the approaches reviewed in the literature have been applied to a dataset from aerial camera arrays. Many of them are not feasible for near real-time processing because of the characteristics of the data collected by aerial camera arrays, including pixel size and illumination, and the consequential challenges in processing the data, including orthorectification and mosaicking.

Persistent Monitoring using Aerial Camera Arrays

Airborne video imaging using camera arrays is a newly evolving technology that enables persistent coverage of a large geographical region depending on the altitude of the platform and the configuration of the array (Palaniappan et al. 2010). The aerial camera array combined with computer vision techniques provides a virtual view of the region being monitored. The surveillance system typically uses a circular flight path to cruise at a constant altitude. As the airplane moves, the camera array is adjusted continuously to maintain the orientation of the camera array fixed around a point on the ground. The geographic region of coverage of the array can remain constant for a number of hours depending on the flight time of the airplane.

A number of potential applications of this system exist, but currently the most widely used applications are law enforcement and event security. In these applications, vehicles or individuals of interest are manually tracked, and the concerned information is relayed to law enforcement personnel on the ground. Automated tracking of vehicles using the data captured by aerial camera arrays has not yet been accomplished. Palaniappan et al. (2010)

found that while the potential applications of automated processing of aerial camera array data are promising, a number of challenges exist. These include the need for improved camera calibration, better estimation of platform dynamics, accounting for lighting variability and seamless image mosaicking. Additionally, many of the existing approaches discussed in the previous section are not capable of processing aerial camera array data because they require higher resolution imagery to analyze the pixels in the neighborhood around the vehicles in static images; whereas the resolution of aerial imagery (when collected at high altitude) is too poor to be able to distinguish vehicles from other objects in the static images. Other approaches use either background subtraction or frame differencing, which cannot be applied to aerial camera array video unless the frames are completely stabilized.

APPROACH AND METHODOLOGY

Our research approach and methodology focuses on implementing and testing image processing and computer vision techniques to explore feasble approaches and improve their performance to overcome challenges with automatically extracting traffic data from a persistent monitoring dataset. The persistent monitoring dataset used in our research is provided by Persistent Surveillance Systems (PSS). PSS utilized a HawkEye II camera array system to collect the imagery. In our research, we have identified numerous challenges with the video sequence that need to be overcome for automated processing to be possible. These challenges are identified as follows:

1. The sub-images from different cameras in the array are not exactly aligned, and the stitched images are seamed, as shown in Fig. 2 (b).
2. The images are not completely stabilized, as shown in Fig. 2 (a) and (b). Notice how the image in Fig. 2 (b) is shifted to the left from Fig. 2 (a).
3. The illumination of sub-images from different cameras in the array is not consistent, as shown in Fig. 2 (d).
4. Even the illumination in single sub-image from a camera in the array is not consistent, as shown in Fig. 2 (d).
5. The video was preprocessed by the monitoring system. The preprocessing software is proprietary and we do not have access to the details of how the image data was preprocessed.
6. The resolution is low, about 0.5 m by 0.5 m per pixel, thus there are few detailed features available to detect vehicles from a static image. As shown in Fig. 2 (c), it is difficult to distinguish cars (right two) and other objects (left two).
7. The images are large, 16384×16384 pixels for 8.05 km by 8.05 km (5 miles by 5 miles). Some image processing and computer vision toolkits cannot support data of such high resolution.
8. The frame rate is only 1 Hertz. On a freeway, a vehicle traveling at 97 km/h (60 mph) travels 26.8m in a second (53.6 pixels), making correspondence between frames very challenging.
9. The amount of data is huge. A single compressed frame is 40-50 MB and a single uncompressed frame is nearly 1 GB, thus our algorithms require high computational and memory efficiency to achieve near real-time execution.

By overcoming these challenges, automated persistent traffic monitoring can be achieved, and diverse traffic data can be extracted with adequate algorithms of vehicle detection and tracking.

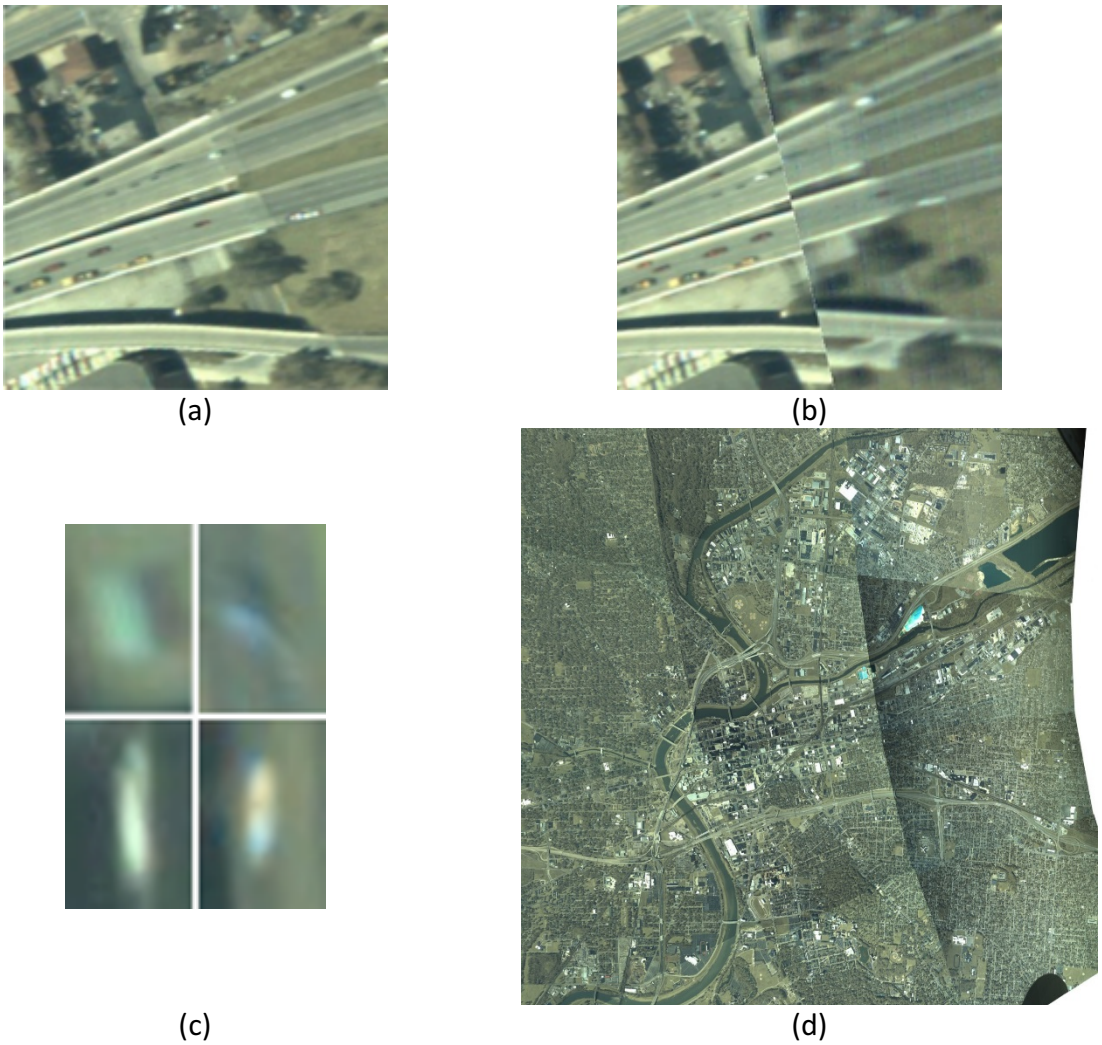


Fig. 2. (a) A section of an image frame; (b) The same section in the next frame, showing a seam; (c) Zoomed portions of a frame showing vehicles (right two) and other objects (left two); (d) A frame of the input data.

To explore the possibility of vehicle tracking for high-resolution aerial imagery, we tried several different heuristic and machine learning approaches including a simple pixel based tracker leveraging information from OpenStreetMap (www.openstreetmap.org). In the following sections, we present and combine the two most promising approaches. The first is a feature-based vehicle tracking framework, whereas the second is based on deep learning. The testing was done on selected vehicles; however, both approaches are scalable to the entire image mosaic.

A Feature-Based Vehicle Tracking Framework

The purpose of the approach presented in this section is to develop a general framework for vehicle tracking with a variety of existing algorithms in digital image processing and computer vision.

This vehicle tracking framework depends on feature detection and matching. It is a heuristic approach based on the fact that the cluster of features representing the tracked object does not significantly morph appearance between consecutive frames. Thus, the sample object can be identified in adjacent frames by explicitly matching the specific representative features of the object with adequate methods (detection, matching and filtering). The basic procedure is as follows:

1. Specify the vehicle to track in the initial frame.
2. For each pair of consecutive frames:
    a. Create ROIs (regions of interest) in both frames.
    b. Extract background information in both ROIs.
    c. Detect (SURF) features in both ROIs.
    d. Identify the cluster of the features representing the tracked vehicle in the 1st frame.
    e. Extract descriptors for the representative features in both frames.
    f. Match descriptors between two ROIs.
    g. Filter matches based on background information and vehicle data.
    h. Identify the cluster of the features representing the tracked vehicle in the 2nd frame.
    i. Collect vehicle information and update the vehicle data.
    j. Predict and update ROIs for the next frame pair.
3. Output vehicle data.

The implementation and results of the framework discussed in this section is coded in Visual Studio 2013 using OpenCV. The results illustrate that the approach is capable of tracking a specified vehicle in the image sequence captured by aerial camera arrays. The performance of this approach depends on the detailed methods applied in this framework.

The results illustrated in Fig. 3 (a) and (b) are based on a SURF (Speeded Up Robust Features) detector and descriptor. A simplified constant acceleration model is also employed in this implementation to estimate speed, acceleration and orientation for the purpose of predicting a vehicle's future location and reducing the feature search and matching range. As shown in Fig. 3 (a), the black car (marked by a magenta circle) is tracked in the traffic flow on a bridge. In Fig. 3 (b), the white truck is also successfully tracked. The performance of this specific application depends on a number of factors, including the contrast of vehicles and disruptors, among which the occlusion and seams are still challenging. Using SURF, this approach is capable of tracking most vehicles with stable appearance in many conditions.

(a)

(b)

Fig. 3. (a) Tracking a black car (8 adjacent frames are shown); (b) Tracking a white truck.

Deep Learning Based Vehicle Detector

Our deep learning based vehicle detector uses a customized deep learning neural network with the Caffe Library developed at University of California at Berkeley. The Caffe Library allows for the customized network to be trained and run on a GPU, speeding up processing.

The deep learning network is designed to detect whether an image patch has a vehicle or not. Fig. 4 (a) shows the detection operation of the network on a single image patch. The image patch is processed by two stages of convolutional / max pooling layers (represented in red) and then two inner product layers (represented in blue). The input image patches are 60 × 60 pixels. The kernel size in the first convolutional layer is 7 × 7 with 20 output feature maps and the kernel size in the second convolutional layer is 5 × 5 with 50 output feature maps. The two max-pool layers both have a kernel size of 3 × 3; however, the stride of the first layer was 1 while the stride of the second layer was 2. For the inner product layers, the first has an output size of 500 while the second has an output size of 2 (one output that represents how likely it is a vehicle, the other that represents how likely it is not a vehicle).

In our dataset, we labeled 1002 points with vehicles and 604 points without vehicles. Image patches were collected from each point. Nine other image patches were collected from each point using rotated versions of the dataset. These patches (16,060 in total) were used to train the network. The rotated patches were collected to help the network learn rotational invariance. The training of the network took about 1 hour on an NVIDIA Tesla K40 GPU. Fig. 4 (b) shows the vehicle detector run on a whole image by splitting the image into many image patches, each of which was run through the detector. Patches were pulled from the image centered about each pixel and tested using the detector. This produced a score for each pixel indicating how likely that pixel contains a vehicle. In Fig. 4 (b), this is represented by a grayscale image, where the brightness indicates the likelihood of a vehicle present. Non-maximal suppression finds the peak in this grayscale image and these peaks are identified as vehicles. For the example input image within Fig. 4 (b), all eight vehicles were detected, three of which were false positives. These false positives can be removed in the tracking steps.

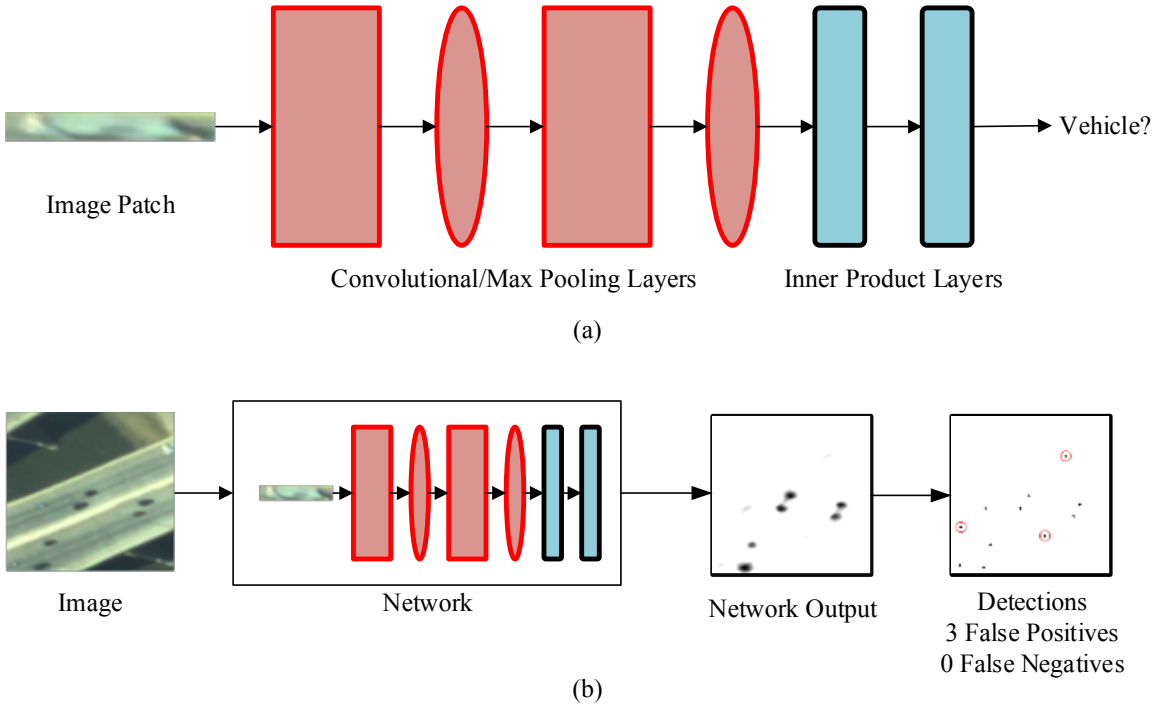


Fig. 4. (a) Diagram of the deep learning detector operating on an image patch; (b) Diagram of the detector being used.

The deep learning network can also be used to match vehicle detections between frames. The Caffe library allows the training of Siamese Networks where two image patches can be sent through an identical set of layers and then compared. This allows us to train a network that can distinguish between different vehicles. The output from this network would be a number representing how likely the two image patches represent the same vehicle.

To test this vehicle detector, the following tracking system was implemented:

1. The vehicle to track is manually selected in the first two frames.
2. Initial vehicle speed and heading are calculated.
3. For each frame:
    a. The detector is used to detect vehicles within 30 pixels of the location predicted by a constant velocity model.
    b. Each detection is compared with the detection in the previous frame using the Siamese Network.
    c. The detection with the highest score is considered the vehicle's position in this frame.
    d. The speed and heading are updated.

Even though the tracking algorithm was relatively simple, the detector was robust enough to provide good results. Fig. 5 shows two vehicles being tracked across image seams and lighting changes. Each image in Fig. 5 represents a frame. There is a close up of one

vehicle in the top left hand corner and of the second vehicle in the bottom right hand corner. Note that vehicles travel nearly 30 m (100 feet) between frames at freeway speeds.

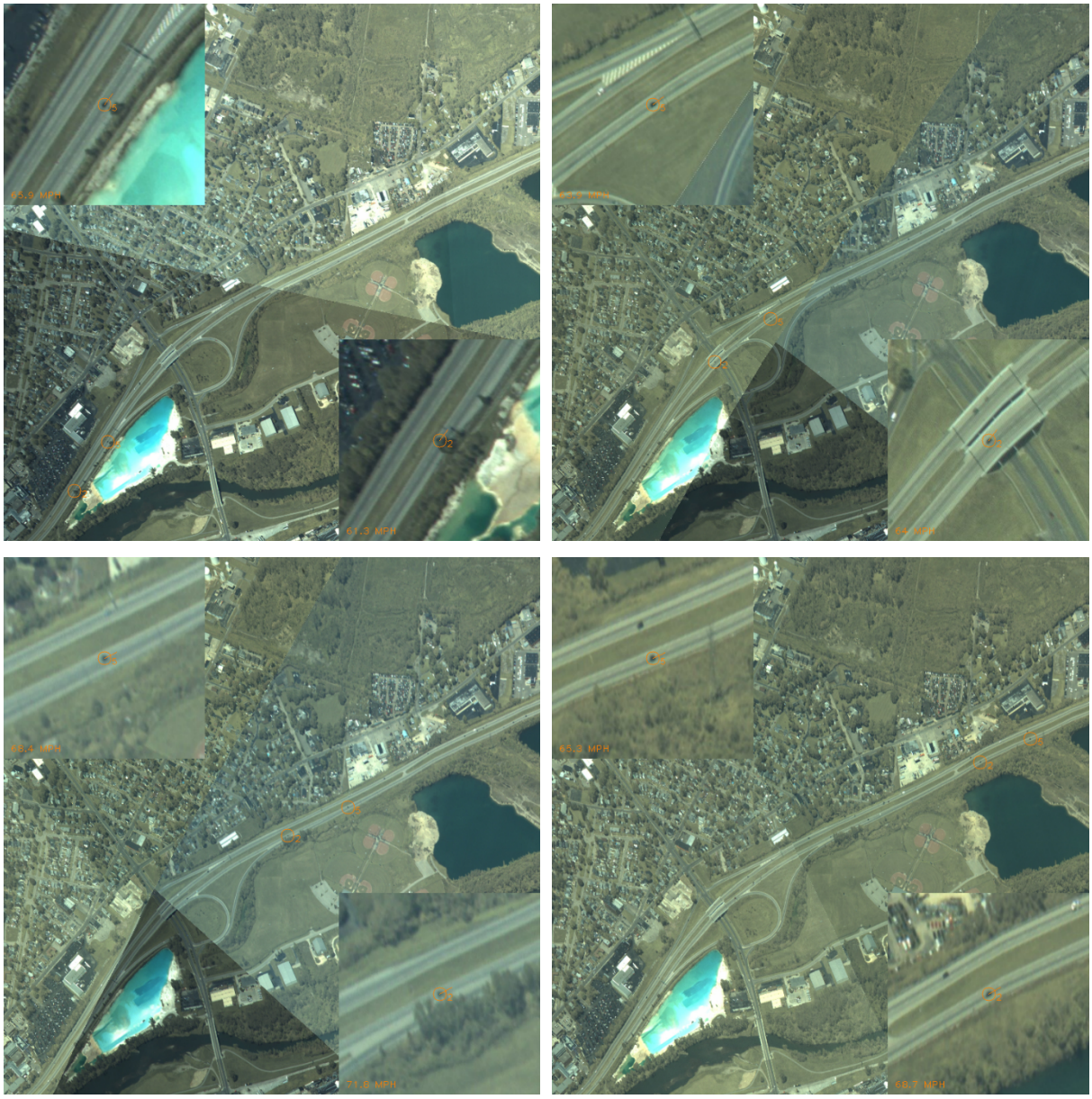


Fig. 5. The deep learning detector tracking two vehicles across the seams and lighting changes.

Data Extraction and Testing

The experimental development of our vehicle tracking approaches and their performance illustrate the feasibility to develop a useful traffic surveillance system to collect traffic data based on aerial camera arrays. Our initial testing showed that our deep learning approach is most promising and potentially capable of achieving expected performance. Due to the quality of the data, vehicle detection for individual frames independently of other

frames is not possible even with extensive training. We combined the two approaches to get the best detection and tracking abilities. The approach we used to collect traffic parameters such as speed, density, and volume included the deep learning vehicle detector to locate potential vehicles, and the feature based tracking to match the vehicles between frames.

The combined system is able to collect the most common traffic parameters measured by traffic monitoring devices, such as speed and volume. An additional parameter that can be directly collected using camera array video is traffic density. This is possible because detection is performed along an entire segment rather than at discrete locations. The following sections discuss the approach used to collect this data from the camera array video.

Speed

The approach for collecting speed data is based on the deep learning test procedure discussed in the previous section except that vehicles are automatically detected rather than manually identified. Fig. 6 illustrates sample speed and acceleration data of a tracked vehicle. Fig. 6 shows how the data are sensitive to the instability of the image frames. Filtering methods can be used to interpolate between frames as well as help neutralize high-frequency vibration of frames. Improved preprocessing to better stabilize image frames combined with more extensive calibration and registration of the image frames will improve the data with less need for filtering; however, this will significantly increase processing time. Further, the filtered speed data only varies by less than 8 km/h (5 mph) throughout the sequence. Filtered acceleration data is even more sensitive to the instability of the image frames.

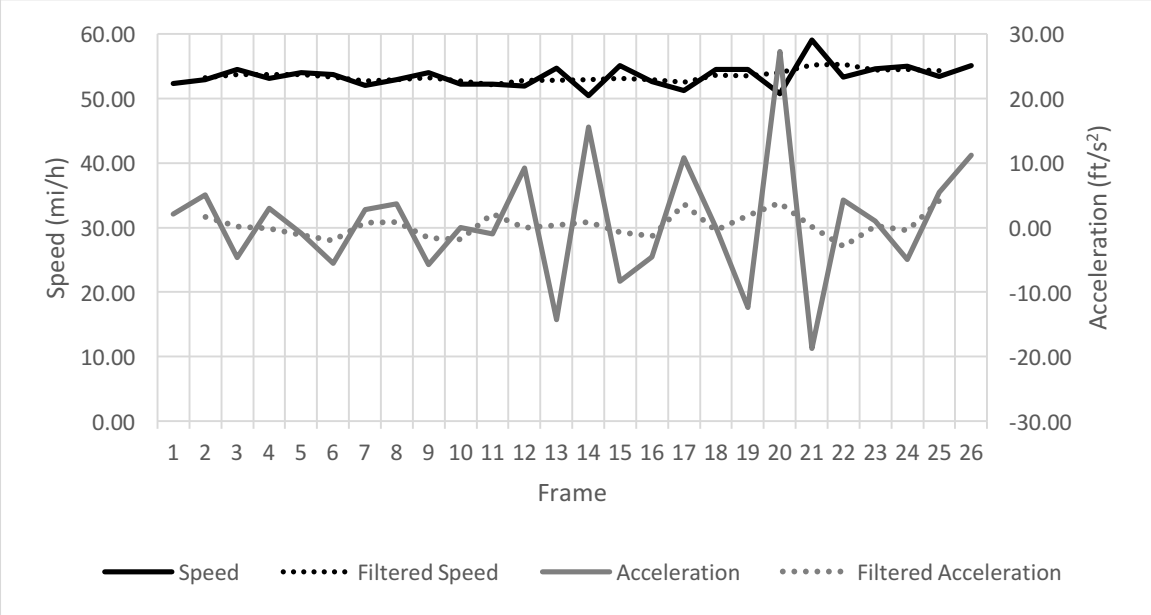


Fig. 6. Sample speed and acceleration data of a randomly selected truck

Density

It is now possible to get near real-time speed data from the millions of real-time anonymous mobile phones, connected cars, trucks, delivery vans, and other fleet vehicles equipped with GPS locator devices. While this speed data is very useful, a better indicator of a facility's

congestion is traffic density. Even when speeds indicate traffic is flowing well, a facility can be close to capacity in which traffic conditions are able to degrade with little notice. Density gives a better indication of overall congestion of uninterrupted flow facilities, which explains its choice as the measure of effectiveness for determining the level of service (LOS) for freeways and multilane highways (Highway Capacity Manual 2010).

Our approach for collecting density data requires an image mask where selected segments are buffered and attributed. Underlying vehicles are detected in these masks using the deep learning approach, and density is calculated for each frame in an image sequence. The training algorithm is capable of identifying a large proportion of vehicles in a single frame; however, the quality and resolution of the image make it virtually impossible to correctly identify all vehicles in a segment. Using feature tracking combined with deep learning applied to a sequence of images, false positives and false negatives can be reduced. An example of this is shown in Fig. 7. The top frame is analyzed with all the detections from the deep learning detector shown as circles. SURF features are tracked between that frame and the next frame to find correspondences, shown in Fig. 7 as black lines. Not all the vehicle detections were able to be matched with the SURF features (blue circles). Detections that do not move are then labeled as false positives (shown as red in Fig. 7). For each detection that did move (green circles), speeds were estimated. These can be used to estimate the average speed over the segment which is useful for calculating volume. For the small segment in Fig. 7, the deep learning detector had 12 detections; of which 6 were tracked with speed estimates and 4 were stationary and determined to be false positives. The vehicle count would be given as 8 which can then be used to calculate the vehicle density for this road segment. The speed for the segment would be the average of the green detections. Applying filtering techniques, an average density can be found for each segment over a predefined period (e.g. 1 minute, 5 minutes, etc.).

Fig. 7 Detections from a frame (top) are filtered using the next frame (bottom).

Volume

Once average speed and density are determined, volume for a segment can be determined by multiplying density by speed. A common approach for determining traffic volumes passing a point is to use a detection device such as an inductive loop detector that can count vehicles that pass over it. Similarly, machine vision sensors use cameras mounted on the side of the road, and vehicles are counted as they pass over a predefined virtual detector. There are multiple problems with this approach if applied to camera arrays. First, image instability from the camera array will require that the detector's location be recalibrated on a continuous basis. Further, because of the 1 Hertz frame rate, vehicles will skip a detector unless the detector is very long. However, a long detector can be touched by multiple vehicles simultaneously if they are closely spaced, resulting in undercounting. Furthermore, because of poor image resolution, using a single image to identify vehicles passing over a virtual detector will lead to errors due to closely space vehicles, shadows, and poor detection results due to the large pixel size resulting in unidentifiable features in many instances.

TABLE 1 Traffic Data Measurements

Road Segments	No. of Lanes	Length (mi)	Count (veh)		Speed (mph)		Density (v/mi/ln)		Volume (v/hr/ln)		LOS		Accuracy	
			A[1]	M[2]	A	M	A	M	A	M	A	M	Density	Volume
OH-4 (WB)	2	3.05	30	36	66.9	69.8	4.9	5.9	329	412	A	A	83.33%	79.82%
OH-4 (EB)	2	3.04	33	36	62.8	66.8	5.4	5.9	341	396	A	A	91.67%	86.19%
I75 (SB)	3	1.68	96	104	63.1	56.2	19.0	20.6	1202	1160	C	C	92.31%	96.42%
I75 (NB)	3	1.68	60	76	58.8	68.2	11.9	15.1	700	1029	B	B	78.95%	68.06%
US-35 (EB) [1]	4	0.83	23	30	57.3	50.0	6.9	9.0	397	452	A	A	76.67%	87.88%
US-35 (WB) [1]	4	0.83	11	12	56.0	65.2	3.3	3.6	186	236	A	A	91.67%	78.94%
US-35 (WB) [2]	3	0.88	54	54	61.5	65.7	20.5	20.5	1258	1345	C	C	100.00%	93.55%
US-35 (EB) [2]	3	0.88	61	67	58.8	62.9	23.1	25.4	1359	1596	C	C	91.04%	85.15%
OH-4 (WB) *	2	3.05	34.66	36.98	65.6	NA	5.7	6.1	373	NA	A	A	93.73%	NA
OH-4 (EB) *	2	3.04	32.12	32.72	62.1	NA	5.3	5.4	327	NA	A	A	98.17%	NA

1 Automatic measurement

2 Manually measured ground truth

* Data based on 50 frames.

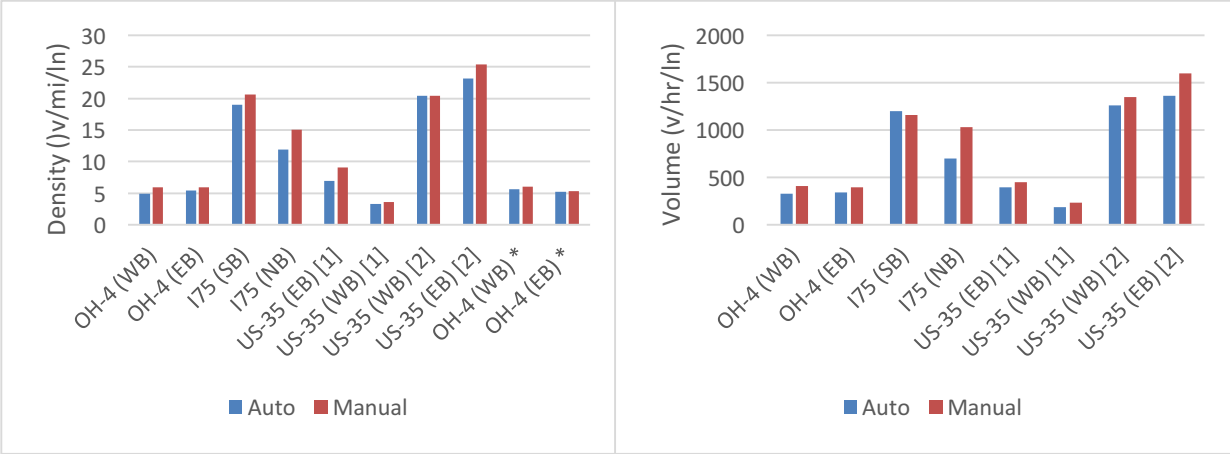
Evaluation and Experimental Results

The automatically and manually collected data for uninterrupted flow segments is shown in Table 1 for all eight segments for a single frame. These segments are from the same imagery with which the network was trained; however only 23 of the 3828 vehicles in the test segments (0.6%) were used in training. Unfortunately, with a limited dataset and randomly selected training samples, some overlap between training and testing was inevitable. Such a small overlap should not detract from the demonstration of the overall reliability and capability of the deep learning detection with feature-based tracking algorithm. The ground truth density was obtained by manually counting vehicles in each frame and the ground truth speed was obtained by averaging the Euclidean distance of the movement across frames of

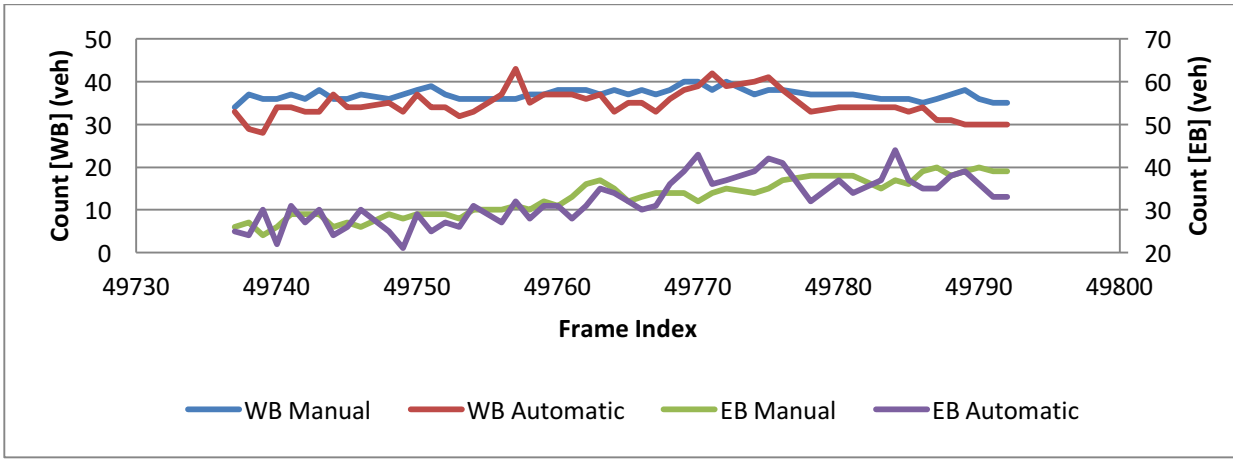
randomly selected vehicles, and volume was obtained by multiplying density and speed. For validation, manually labeled ground truth data is compared to the automatic measurements. We found the density data presents high accuracy while the speed data can also be accurate based on the aforementioned discussion, both of which can lead to accurate estimates of LOS and volume data.

The automatic density data is precise. Ground truth densities were determined by meticulously hand counting vehicles along the segments for each frame for the entire one minute video and taking the average. With only one frame, the average errors of estimates of density are 11.8% for all eight segments. However, when the density measurements are averaged across 50 frames, the accuracy is as high as 98.2%, as shown in Fig. 8 (a). This result indicates that the proposed approach is reliable for collecting density data from a sequence of camera array images. From density we can calculate the LOS, which due to the reliable density estimates, were 100% accurate for the segments evaluated.

The speed data is very sensitive to the instability of the imagery data. Ground truth speeds were calculated by randomly sampling a subset of vehicles for each segment and manually tracking them over the images and averaging their speeds. The instantaneous speed based on the correspondence of two adjacent frames oftentimes does not give a useful result because of the shifting and rotation of the second frame. However, the average speed across multiple frames can be very accurate and useful with filter techniques.



(a) (b)



(c)

Fig. 8. Automatic vs. manually collected ground truth for (a) density, (b) volume, and (c) vehicle counts over 50 seconds on OH-4.

The volume data was found to be accurate. Even though the instantaneous speed is not reliable, the volume data calculated with it still presents reasonable results as shown in Fig. 8 (b). The average accuracy is 84.5% for all eight segments. The accuracy of volume data can be greatly improved based on filtered average speed across multiple frames.

Tracking across multiple frames produces reliable data. The performance of the proposed approach is powerful in counting vehicles, leading to precise average density over time as shown in Fig. 8 (c). Similarly, other measurements are also precise if averaged across multiple frames. Unfortunately, it is not practical to manually extract speed for each vehicle for every pair of frames for the entire sequence to collect precise speed ground truth for validating this approach.

Many of the challenges due to the aerial camera array aforementioned. have been addressed. Our approaches are robust enough to deal with the instability, seams, and inconsistent illumination of the images with limited effects on the accuracy of the collected traffic parameters. Future work will significantly improve the accuracy with more sophisticated algorithms. The resolution, image size and frame rate are the special features of our research which can be handled effectively by our approaches. One remaining challenge is the computation time: It takes 640 seconds on average to process a 64.80 km^2 (25 mi^2) frame which means 9.87 seconds for a 1 km^2 region (25.6 second for a 1 mi^2 region). More work is needed to make the computation efficient for real-time processing.

FINDINGS; CONCLUSIONS; RECOMMENDATIONS

We have proposed a novel automated traffic surveillance system based on aerial imagery from camera arrays that are not conducive to previous tracking methods. Instead, due to the unique issues present in these data, we tested two new ideas based on variations of previous tracking methods with promising results. The first was to develop a general framework for vehicle tracking based feature detection; the second was a vehicle detection approach based

on deep learning. The combination of them was found to be the most promising in our testing and thus was used to collect speed, density, and volume for uninterrupted flow segments throughout the 64.80 km^2 (25 mi^2) coverage area.

The evaluation was conducted for facilities of different levels of congestion as indicated by the different LOS. Quantitative evaluation showed that our current deep leaning based detector combined with feature-based tracking accurately provides pertinent traffic data over a 64.80 km^2 (25 mi^2) area, even with a very difficult dataset. The proposed approach is able to provide single frame estimates of density with an average accuracy of 88.2%. When estimating density over 50 frames, the accuracy is over 93%. The LOS estimates are 100% accurate when measured for the uninterrupted facilities, and the approach is able to provide reasonable estimates for average vehicle speed and traffic volume.

The primary research objective has been to investigate ways to process aerial camera array imagery to extract valid and useful macroscopic traffic data for diverse applications. The evaluation has shown that the proposed system is capable of collecting speed, density, and volume data with an acceptable level of accuracy for many applications. With further research, improved video preprocessing, enhanced resolution, and a higher frame rate, the accuracy of tracking vehicles can be improved significantly which will eventually allow the envisioned system to be able to accurately map the location of vehicles throughout a camera array image sequence. By mapping the locations of vehicles in a spatio-temporal manner, additional traffic parameters can be extracted microscopically. These may include turning movements at intersections, traffic signal timings, identification of erratic drivers; and origin-destination data. Further, such a system could facilitate traffic management through traffic data mining. Traffic data mining methods provide a means to generate dependable patterns from large traffic data sets with less complexity than current approaches. When traffic predictions are accurate, more efficient management of the traffic network is possible. A digital real-time “traffic map” created by the envisioned system will provide a robust data set where data mining methods could be applied to enhance traffic management and provide data for a variety of traffic studies.

REFERENCES

Angel, A., Hickman, M., Chandnani, D., and Mirchandani, P. (2002). "Application of aerial video for traffic flow monitoring and management." *Proc., 7th Int. Conf. Appl. Adv. Technol. Transp.,* ASCE, Reston, VA, 346-353.

Cao, X., Lan, J., Yan, P., and Li, X. (2012). "Vehicle detection and tracking in airborne videos by multi-motion layer analysis." *Mach. Vision. Appl.,* 23(5), 921-935.

Chen, B., and Medioni, G. (2015). "3-D mediated detection and tracking in wide area aerial surveillance." *Proc., 2015 IEEE Winter Conf. Appl. Comput. Vis. (WACV),* IEEE Computer Society, Los Alamitos, CA, 396-403.

Cheng, H., Weng, C., and Chen, Y. (2012). "Vehicle detection in aerial surveillance using dynamic Bayesian networks." *IEEE Trans. Image Process,* 21(4), 2152-2159.

Du, X., and Hickman, M. (2012). "Estimation of a road mask to improve vehicle detection and tracking in airborne imagery." *Transp. Res. Rec.,* (2291), 93-101.

Hickman, M. D., and Mirchandani, P. B. (2006). "Uses of airborne imagery for microscopic traffic analysis." *Proc., 9th Int. Conf. Appl. Adv. Technol. Transp.,* ASCE, Reston, VA, 238-243.

Highway Capacity Manual. (2010). Transportation Research Board, Washington, DC.

Hinz, S. (2004). " Detection of vehicles and vehicle queues for road monitoring using high resolution aerial images." *Proc., 9th World Multi-Conf. Syst. Cybern. Informatics (WMSCI),* International Social Science Council, Paris, France, 405-410.

Kalal, Z., Mikolajczyk, K., and Matas, J. (2012). "Tracking-learning-detection." *IEEE Trans. Pattern Anal. Mach. Intell.,* 34(7), 1409-1422.

Kim, Z., and Malik, J. (2003). "Fast vehicle detection with probabilistic feature grouping and its application to vehicle tracking." *Proc., 9th IEEE Int. Conf. Computer Vision (ICCV),* IEEE Computer Society, Los Alamitos, CA, 524-531.

Moon, H., Chellappa, R., and Rosenfeld, A. (2002). "Performance analysis of a simple vehicle detection algorithm." *Image Vis. Comput.,* 20(1), 1-13.

Nguyen, T. T., Grabner, H., Bischof, H., and Gruber, B. (2007). "On-line boosting for car detection from aerial images." *Proc., 2007 IEEE Int. Conf. Res. Innov. Vis. Future (RIVF),* IEEE Computer Society, Los Alamitos, CA, 87-95.

Palaniappan, K., Bunyak, F., Kumar, P., Ersoy, I., Jaeger, S., Ganguli, K., Haridas, A., Fraser, J., Rao, R. M., and Seetharaman, G. (2010). "Efficient feature extraction and

likelihood fusion for vehicle tracking in low frame rate airborne video." *Proc., 13th Int. Conf. Inf. Fusion,* IEEE Computer Society, Los Alamitos, CA, 1-8.

Palaniappan, K., Rao, R. M., and Seetharaman, G. (2011). "Wide-area persistent airborne video: architecture and challenges." *Distributed Video Sensor Networks,* Springer, 349-371.

Pelapur, R., Candemir, S., Bunyak, F., Poostchi, M., Seetharaman, G., and Palaniappan, K. (2012). "Persistent target tracking using likelihood fusion in wide-area and full motion video sequences." *Proc., 15th Int. Conf. Inf. Fusion,* IEEE Computer Society, Los Alamitos, CA, 2420-2427.

Prokaj, J., and Medioni, G. (2014). "Persistent tracking for wide area aerial surveillance." *Proc., 27th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* IEEE Computer Society, Los Alamitos, CA 1186-1193.

Reinartz, P., Lachaise, M., Schmeer, E., Krauss, T., and Runge, H. (2006). "Traffic monitoring with serial images from airborne cameras." *ISPRS J. Photogramm. Remote Sens.,* 61(3), 149-158.

Saleemi, I., and Shah, M. (2013). "Multiframe many–many point correspondence for vehicle tracking in high density wide area aerial videos." *Int. J. Comput. Vis.,* 104(2), 198-219.

Saunier, N., and Sayed, T. (2007). "Automated analysis of road safety with video data." *Transp. Res. Rec.,* (2019), 57-64.

Shi, X., Li, P., Ling, H., Hu, W., and Blasch, E. (2013). "Using maximum consistency context for multiple target association in wide area traffic scenes." *Proc., 38th IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP),* IEEE, Piscataway, NJ, 2188-2192.

Tsai, Y., Wang, C., and Wu, Y. (2011). "A vision-based approach to study driver behavior in work zone areas." *Proc., 3rd Int. Conf. Road Safety Simulation,* TRB, Washington, DC, 14-16.

Tuermer, S., Leitloff, J., Reinartz, P., and Stilla, U. (2010). "Automatic vehicle detection in aerial image sequences of urban areas using 3D HoG features." *Proc., ISPRS Technical Commission III Symposium Photogramm. Comput. Vis. Image Analysis (PCV),* International Society for Photogrammetry and Remote Sensing, Saint-Mande, France, 50-54.

Xiao, J., Cheng, H., Sawhney, H., and Han, F. (2010). "Vehicle detection and tracking in wide field-of-view aerial video." *Proc., 23rd IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* IEEE Computer Society, Los Alamitos, CA, 679-684.

Zhao, T., and Nevatia, R. (2003). "Car detection in low resolution aerial images." *Image Vis. Comput.,* 21(8), 693-703.

APPENDIX

Publications, presentations, posters resulting from this project:

1. Zhao, X., D. Dawson, W. Sarasua, and S. Birchfield, "Automated Traffic Surveillance from an Aerial Camera Array," August, 2014, **White paper.**
2. Sarasua, W, "Automated Traffic Surveillance from an Aerial Camera Array" Presented at the STC Research Conference, Southeastern Region, Southeastern Transportation Center, Knoxville, TN, Sept, 2014. **Presentation.**
3. Zhao, X., D. Dawson, W. Sarasua, and S. Birchfield, "An Automated Traffic Surveillance System with Aerial Camera Arrays: Data Collection with Vehicle Tracking," Presented at Transportation Research Board 95th Annual Meeting 16-6783), January, 2016 **Publication and Poster**
4. Zhao, X., D. Douglas, W. Sarasua, and S. Birchfield. An Automated Traffic Surveillance System with Aerial Camera Arrays: Data Collection with Vehicle Tracking, 2016 UTC Conference for the Southeastern Region, Southeastern Transportation Center, Knoxville, TN, April, 2016. **Poster**
5. Zhao, X., D. Dawson, W. Sarasua, and S. Birchfield, "An Automated Traffic Surveillance System with Aerial Camera Arrays: Data Collection with Vehicle Tracking," Journal of Computing in Civil Engineering, ASCE, July, 2016 **Publication** (Note: the paper was just revised based on reviewer comments and is going through editorial review prior to publication)

   Two other papers two Ph.D. dissertations are in progress. The researchers also plan to pursue a patent.