A Framework for Realistic Vehicular Network Modeling using Planet-scale Public Webcams

Gautam S. Thakur
CISE, University of Florida
Gainesville, USA
gsthakur@cise.ufl.edu

Pan Hui
Telekom Innovation
Laboratories, Berlin, Germany
pan.hui@telekom.de

Ahmed Helmy
CISE, University of Florida
Gainesville, USA
helmy@cise.ufl.edu

ABSTRACT

Realistic design and evaluation of vehicular mobility has been particularly challenging due to a lack of large-scale real-world measurements in the research community. Current mobility models and simulators rely on artificial scenarios and use small and biased samples. To overcome these challenges, we introduce a novel framework for large-scale monitoring, analysis, modeling, and visualization of vehicular traffic using freely available online webcams. We follow a data-driven approach that examine six metropolitan regions' more than 800 locations and 25 million vehicular mobility records around the world. Initial analysis of traffic densities show 80% temporal correlation during various hours of a day. The modeling of empirical traffic densities against known theoretical models show less than 5% deviation for heavy-tailed distributions such as Weibull. We believe this framework and the dataset provide a much-needed contribution to the research community for *realistic* and *data-driven* design and evaluation of vehicular networks.

Categories and Subject Descriptors

H.2.8 [[**Database Management**]]: Database Applications-Data mining, Image databases, Spatial databases and GIS.

General Terms

Experimentation, Human Factors, Measurement.

Keywords

Urban Infrastructure, Vehicular Traffic, Social Networks.

1. INTRODUCTION

Research in the area of vehicular networks has increased dramatically in the recent years. With the proliferation of mobile networking technologies and their integration with the automobile industry, various forms of vehicular networks are being realized. These networks include vehicle-to-vehicle[4], vehicle-to-roadside[10], and vehicle-to-roadside-to-vehicle architectures. Realistic modeling, simulation and informed design of such networks face several challenges, mainly due to the lack of large-scale community-wide libraries of vehicular data measurement, and representative models of vehicular mobility.

Earlier studies in this area have clearly established a direct link between *vehicular density* distribution and the performance[5, 16] of vehicular networks primitives and mechanisms, including broadcast and geocast protocols[1]. Initial efforts to capture realistic vehicular density distributions were limited by availability of sensed vehicular data[20]. Hence, there is a need to collect and conduct vehicular density modeling using larger scale and more comprehensive datasets. Furthermore, commonly used assumptions, such as exponential distribution[19] of vehicular inter-arrival times[1], have been used to derive many theories and conduct several analyses, the validity of which bears further investigation.

In this study, we provide a novel framework for the systematic monitoring, analysis and modeling of vehicular traffic density distributions at a large-scale. To avoid the limitations of sensed vehicular data, we instead utilize the existing global infrastructure of tens of thousands of video cameras providing a continuous stream of street images from dozens of regions around the world. Millions of images captured from publicly available traffic web cameras are processed using a novel density estimation algorithm to build an extensive measurement library of spatio-temporal vehicular traffic densities. We perform a comprehensive analysis of this data to characterize the underlying statistical patterns at individual intersections and highways of major regions. Briefly, the temporal correlations between consecutive hours of individual locations are nearly 80% correlated, but go down to 30% for a 3-4 hours lag difference. We also investigate the best distribution fitting by comparing the frequencies, observed in the empirical density distribution to the expected frequencies of the theoretical distribution. We discover that empirical values closely follow (less than 3% deviation on KS-test) heavy-tailed models such as 'log-logistic' and 'log-gamma', and Weibull distributions. *These results question the adequacy to apply exponential distribution for vehicular traffic modeling and can impact the design and evaluation of vehicular networks.* The contributions of this work are:

- To the best of our knowledge, we provide by far the largest and most extensive library for future vehicular network analysis. This potentially addresses a severe shortage of such datasets in the community. The li-

HotPlanet'12, June 25, 2012, Low Wood Bay, Lake District, UK.

brary will be made available to the research community in the future.

- A fast algorithm to efficiently process millions of images for novel traffic density estimation.

- Establish heavy-tailed models such as log-logistic and Weibull distributions as the most suitable fits for modeling empirical values of vehicular traffic density.

We believe our framework helps 'fills a gap' between expected and realized necessity for the 'design and evaluation of realistic and data-driven models' for the future generations of vehicular networks.

Next, in section 2, we discuss related work. In section 3, we propose framework and introduce our dataset and explain vehicular traffic density estimation process, perform knowledge discovery, and statistical modeling of empirical traffic time series. In section 4, we briefly discuss impacts and challenges in vehicular networks. Finally, we conclude in section 5 and give insight into future work.

2. RELATED WORK

Large-scale mobility datasets are very important for the mobile networking and computing research communities, but collecting them is even more challenging and usually expensive [4, 9, 17]. In some cases, commercial vendors log number of vehicles, GPS coordinates, speed and movement traces. However, there are three downsides to it. First, these traces are not publicly available to the research community. Second, they contain only particular vehicles with vendor specific hardware. Third, they are from individual vehicles with short driving distances and in most cases with non-repetitive journeys. Invariably, these issues undermine the efficacy of using them for any kind of longitudinal analysis. In this paper, we propose an inexpensive method to collect global scale vehicular mobility traces using thousands of freely available webcams that provide continuous and fine-grained monitoring of the vehicular traffic. Simulation tools like CORSIM[8] and VISSIM[11] are geared to model specific scenarios for planning future traffic conditions on a micro-mobility and small scale level. But they lack tools to perform network analysis[17] . From a networking perspective, mobility models[6, 13] and routing[22] techniques investigate how mobility impact the performance of routing protocols[2]. If the mobility model is unrealistic then routing performance becomes questionable[14]. We need models inspired from real datasets- by way of this work, we believe a comprehensive set of parameters can be extracted to develop such models. In a recent work, Bai et.al [1] analyzed spatio-temporal variations in vehicular traffic for the purpose of inter-vehicle communications. Data collected from realistic scenarios shows the effectiveness of exponential model for highway vehicle traffic. Along the same lines, quantitative characteristics of vehicle arrival patterns on highways is studied in [12]. These findings enrich traffic modeling, but were carried out on very small sample of data and mainly localized to one or two locations. In our study, we use 42 days of vehicular imagery data from six regions to model the traffic and characterize its density distribution.

3. VEHICULAR MOBILITY FRAMEWORK

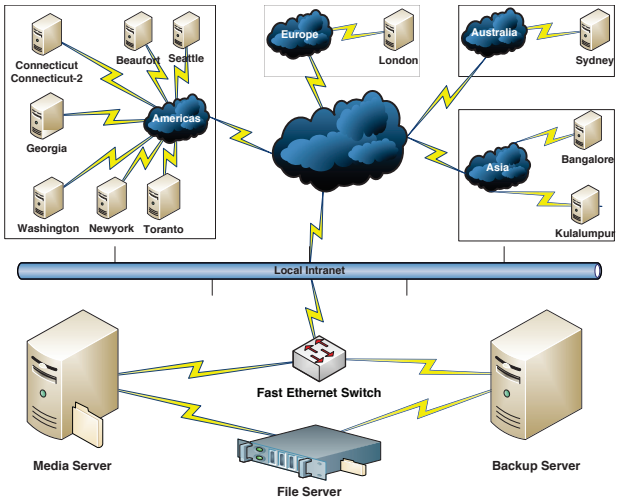


Figure 2: Distributed system architecture for vehicular imagery collection on planet-scale.

In this section, we describe our proposed framework, shown in Figure 1, which is comprised of three parts: (i) Measurements and pre-processing, (ii) Knowledge discovery, and (iii) Modeling and analysis. The measurements and preprocessing component (in green, fully covered) is responsible to capture imagery snapshots, sanitize data and generate a quantifiable value of vehicular traffic, hereafter known as traffic density(d). The knowledge discovery(orange, partially covered) focuses on applying data mining tools to extract traffic patterns, and spatio-temporal information. This activity can help to develop rich mobility scenarios. Next, the modeling and analysis component focus on characterizing the vehicular traffic densities. It can aid in designing and developing new data-driven vehicular mobility models and simulators. Next, we discuss these components.

3.1 Measurement and Vehicular Density Estimation

Table 1 summarizes the dataset used in this research; six regions/cities, the time span of the samples, the sampling rate and the number of camera's/sample locations. On average, we download 15 gigabytes of imagery data per day from over 2,700 traffic cameras, with an overall dataset of 7.5 terabytes containing around 125 million images. In this paper, for a fair comparison, we have selected only six regions with nearly similar time granularity of traffic snap shots, as shown in Table 1. The subset of dataset used here is huge with 25M records. Figure 2 shows the distributed system architecture for vehicular imagery collection on planet-scale. Figure 3 shows a geological snapshot of the cameras deployed in London and Sydney, as an example. The area covered by the cameras in London is 950km^2 while that in Sydney is 1500km^2. Finally, note that since these cameras do not have night vision, we limit our study to the hours between 7am and 6pm.

3.1.1 Background Subtraction

The snapshots taken by every traffic camera (at intervals ranging from 20-60 seconds) first pass a background estimation and subtraction phase. These are then used to estimate

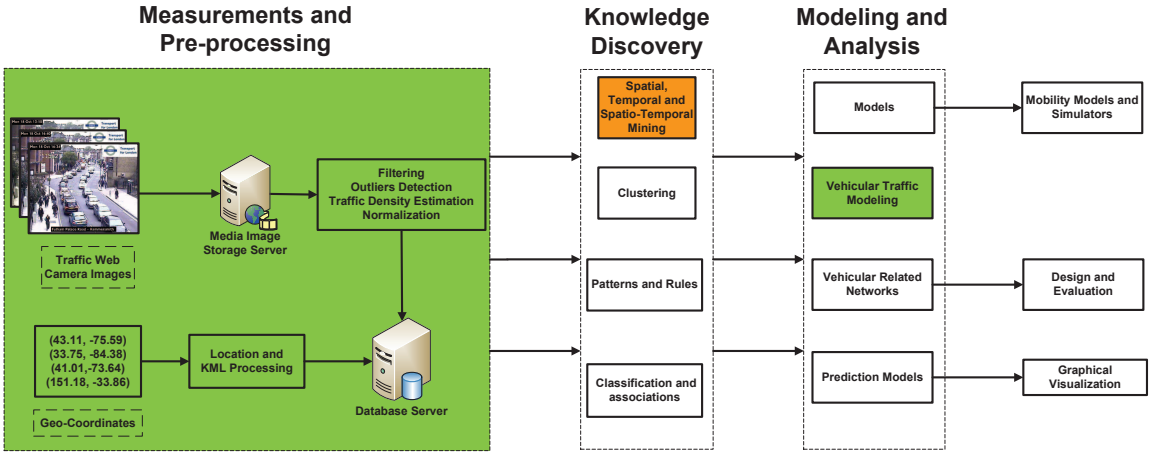


Figure 1: A framework for monitoring, analysis and modeling of vehicular traffic densities.

Region	# of Cameras	Duration	Interval	Records	Database Size	Routes
Connecticut	120	21/Nov/10- 20/Jan/11	20 sec.	7.2 million	435 GB	74,801
London	182	11/Oct/10 - 22/Nov/10	60 sec.	1 million	201 GB	32,580
Seattle	121	30/Nov/10 - 01/Mar/11	60 sec.	8.2 million	600 GB	7,656
Sydney	67	11/Oct/10 - 05/Dec/10	30 sec.	2.0 million	350 GB	4,422
Toronto	89	21/Nov/10 - 20/Jan/11	30 sec.	1.8 million	325 GB	43,055
Washington	240	30/Nov/10 - 01/Mar/11	60 sec.	5 million	400 GB	59,809
Total	**819**	**-**	**-**	**25.2 million**	**2311 GB**	**222323**

Table 1: Global Webcam Dataset

the *traffic density* arriving per unit time as opposed to a car count.

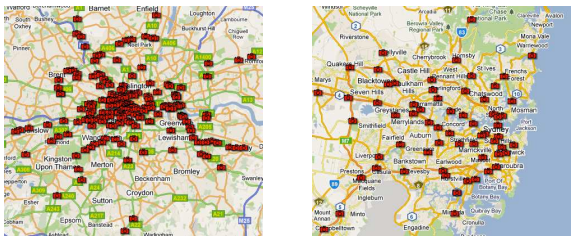
While a car count might seem preferable to a traffic density measure, there are several practical challenges. A car count requires a far greater computational cost due to the effort required to isolate each object. Traffic congestion further complicates matters when cars occlude each other, making it difficult to segregate cars based on edge structures. In addition, vehicles at the far end of the road are small in the image and cannot be detected by these algorithms.[1]

Background subtraction is a standard method for object localization in image sequences with fixed cameras, where the frame rate is lower than the velocity of the objects to be tracked (i.e. cars move out of the scene typically at a rate exceeding 1 minute). The models of background are based on the observation that *background* does not change significantly (in comparison to foreground/objects) across time. Any part of an image that does not fit with that model is deemed as *foreground/object*. These foreground regions are then further processed for the detection of desired objects.

The background model used here assumes that the distribution of background pixel values may be modeled as a weighted sum of Gaussian distributions. Our approach follows closely to those proposed by [3, 15, 18] because of their reliability and robustness to sensitive changes in the lighting conditions. In our approach, the observed pixel value is modeled by a weighted sum of Gaussian kernels. Let x_t represent a pixel value in the t^{th} frame, then the probability of observing this value is assumed to be:

$$p(x_t) = \sum_{i=1}^{K} w_i^t * \mathcal{N}(\mu_{i,t}, \Sigma_{i,t}) \quad (1)$$

where $\mathcal{N}(\mu_{i,t}, \Sigma_{i,t})$ is the i^{th} kernel with mean $\mu_{i,t}$ and covariance matrix $\Sigma_{i,t}$, and w_i^t is the weight applied to that kernel such that $\sum_i w_i^t = 1$. We assume that RGB channels are uncorrelated thus the covariance matrix for each kernel is diagonal.[2] When a new frame arrives, the pixel values are compared to the kernels to determine if it is likely that this value was drawn from a distribution with $\mathcal{N}(\mu_{i,t}, \Sigma_{i,t})$ (us-



(a) London (b) Sydney

Figure 3: Traffic cameras in London and Sydney. The red dots show the location of cameras deployed giving an idea of their distribution in the city.

[1] Another solution could be to only count cars that are close to the camera; while this is definitely an option for video data, for snapshot data it would result in those distant cars having left the scene before the next snapshot; the net effect being that the maximum observed car count at a junction is truncated causing problems in the multivariate analysis later on.

[2] Thus reducing the number of unknown parameters.

Camera	df	$\beta_0(\alpha = 0.95)$	$\beta_1(\alpha = 0.95)$	R^2	p	ρ
1	100	-1.19±0.046	0.03±0.003	0.7922	0	0.91
2	100	-3.25±0.130	0.09±0.007	0.8579	0	0.92
3	100	8.16±0.045	0.10±0.005	0.9308	0	1.00
4	100	8.16±0.045	0.10±0.005	0.9308	0	1.00
5	100	8.16±0.045	0.10±0.005	0.9308	0	1.00
6	100	-2.13±0.112	0.07±0.008	0.7499	0	0.88

Table 2: Summary of regression analysis

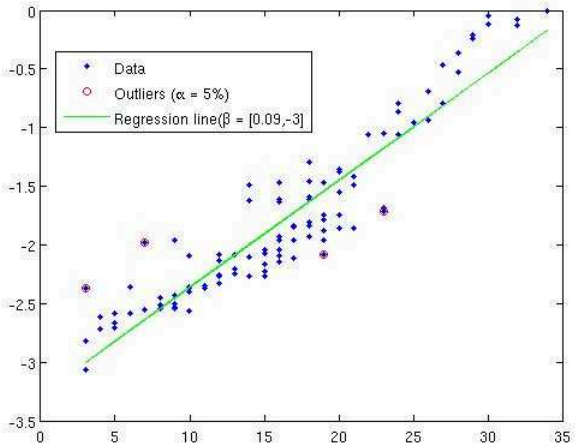


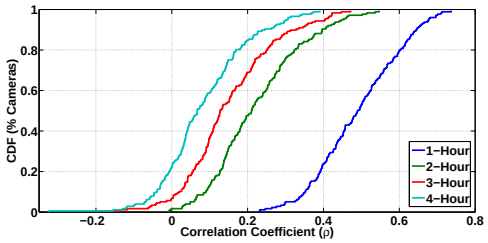
Figure 4: A comparison of traffic densities with number of cars.

ing for example a 95% confidence interval). If so, $\mu_{i,t}$, $\Sigma_{i,t}$ and w_i are updated using exponential filters; if not, a new kernel is created and the existing kernel with the lowest w_i is eliminated (see [18] for specifics). Short lived kernels and their associated pixels are deemed to be possibly foreground producing a binary map. Morphological operations are then applied to this map to remove noise and any blobs with area smaller than a certain threshold.

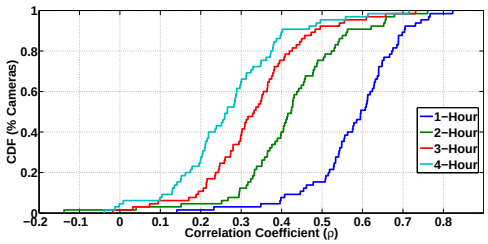
The view of most cameras used in this study is along the direction of the road and this perspective skews the size of objects on an image [7]. To counter this effect, we weigh each foreground pixel with the exponent of it's distance from the bottom of the image. Thus a pixel in the bottom of the image will be weighted less (objects appear larger at the bottom than on the top) than a pixel at the top. While this weighting is not exact and does produce some warping as we shall see in the ground truth validation (next) section; the warping is not excessive given the advantage that the weighting is simple and does not require manually tuning at each camera.

3.1.2 *Ground Truth for Validation*

To test the performance of the car density capture, six cameras were selected at random and 102 images from each were examined by hand to produce a *ground truth* count for the number of cars. This ground truth was then regressed against the measured car density to check if the relationship is linear. The regression from one camera is shown in Figure 4 and shows a reasonable fit. There are some outliers, especially at low levels of traffic and there also appears to be a slight non-linear relationship between the ground truth and measured car density due to the warping effect of perspective (discussed above). Table 2 shows the summary statistics for the regression analysis including Spearman's



(a) London



(b) Sydney

Figure 5: CDF showing correlation of traffic densities between hour differences of the day.

correlation coefficient, ρ, which seems to imply that there is a perfect monotonic non-linear correlation for camera's 3 to 5.[3] Overall, the analysis shows that while there are some errors, the relationship between the actual and measured number of cars is sufficiently clear to allow analysis.

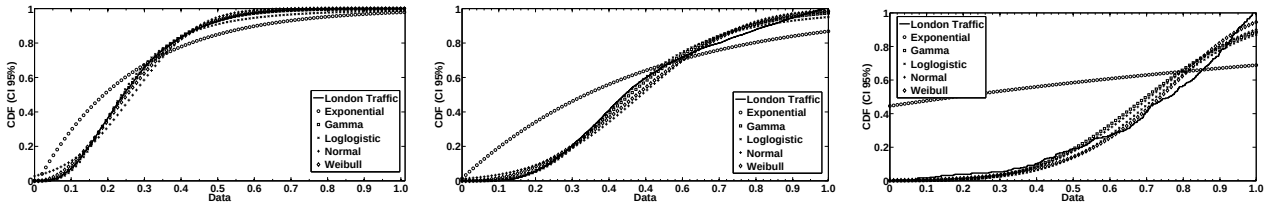
3.2 Knowledge Discovery

We investigate correlation coefficients(ρ) to measure the degree to which traffic of a camera is linearly associated with itself for 42 days. In our case, we are using this to analyze the change in traffic densities. We analyze the correlations for 1-4 hour lags for each camera against itself during 12 hours of the day, from 7 AM to 6 PM. For example, we investigate what the correlation is between the traffic at 7 AM and 8 AM(1-hour lag), 1 PM and 3 PM(2-hour lag) etc. In Figure 5, we show CDF for various hours lag of the day. For the city of Sydney the hourly traffic change is highly correlated, almost 80% of cameras' next hour traffic is 70% correlated to its current hour. For next two hours from the current, the traffic for 80% of the cameras are only 50% or less correlated. And around 60% cameras have only 30% correlation for a time lag of 3-4 hours. While in case of the city of London, the next hour traffic density for 80% cameras is close to 60% correlated to the current hour. It goes further down to 30% for next two hours and around 15-20% for a 3-4 hour difference. *Thus, vehicular traffic has temporal richness, which in-turn affects the mobility of vehicles and therefore, have an impact on the performance of routing protocols[2].* Similar trends are observed in other regions, but omitted here for brevity.

3.3 Modeling and Analysis

The objective of this study is to help understand the underlying statistical patterns. Traffic density is an approxi-

[3]The other notation in Table 2 is standard regression notation: *df* denotes the degrees of freedom. α and β are the regression coefficients as $y = \alpha x + \beta$, R^2 is the % of variance explained, see Equation eqn:r2, p is the p-value.



(a) London(L) (b) London(M) (c) London(H)

Figure 7: CDF plot for three varying traffic intensities, Low(L), Medium(M) and High(H).

Table 3: Dominant distribution as Best Fits[Ranked and % Deviation]

Region	1^{st} Best Fit	2^{nd} Best Fit	3^{rd} Best Fit	$\leqslant$3%	$\leqslant$5%
Connecticut	L[91%]	G[5%]	W[4%]	L[50%], W[2%], G[1%]	L[93%], W[13%], G[10%], E[5%]
London	G[38%]	L[29%]	W[26%]	G[20%], L[15%], W[10%], N[8%]	G[55%], L[51%], W[44%], N[23%]
Sydney	L [70%]	G[17%]	W[14%]	L[65%], G[22%], W[8%],	G[49%], W[37%], N[6%]
Toronto	L[40%]	G[27%]	W[26%]	G[18%], W[17%], L[9%], E[3%]	W[72%], L[69%], G[63%], E[24%], N[1%]
Washington D.C.	L[80%]	W[11%]	G[7%]	L[60%], W[8%], G[6.54%], E[4%]	L[91%], W[35%], G[30%], E[14%]
Seattle	W[36%]	L[34%]	G[29%]	W[16%], G[14%], L[4%]	G[55%], W[47%], L[35%]
E = Exponential, G = Log-gamma, L = Log-logistic, N = Normal, W = Weibull					

(a) $d = 2023, 0.28$ (b) $d = 5400, 0.55$ (c) $d = 9230, 0.93$

Figure 6: Traffic with varying densities[(a)low/(b)medium/(c)high] is shown. The first values is the result of background subtraction and later is the normalized value.

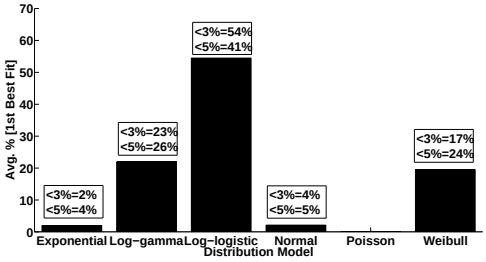


Figure 8: Best fits covering locations of six regions. The values in the box show deviation.

mation of traffic on roads. This assumption is different from counting cars such as by using loop detectors. In Figure 6, we show three traffic scenarios of varying intensities from low to fully congested locations.

Here, we focus on modeling empirical traffic densities against known theoretical distributions. We have filtered these distributions for the purpose of showing the maturity in our studies to select and identify the statistical patterns without much deviations. To ensure the validity of our results, we have also performed several goodness of fit test using Maximum likelihood estimation (MLE) and Kolmogorov-Smirnov test to measure average deviation and compare the values in the density vector to known distribution.

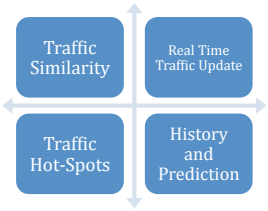


Figure 9: Four visualization scenarios for modeling vehicular traffic density.

Traffic Flow Characterization

We systematically model individual locations' empirical traffic density distribution against well known theoretical ones. In order to match, we use six theoretical distributions: exponential, log-gamma, log-logistic, normal, poisson, and weibull. Figure 7 show CDFs for three individual cameras of London, with low, medium and high traffic densities. We can see that traffic at individual cameras can vary a lot, but in general log-logistic, gamma and weibull distribution can capture some of the key features. We rank these distributions (based on KS-tests) in Table 3, with four out of six regions' individual locations have log-logistic as the 1^{st} best fit, while Toronto has gamma distribution. In Table 3, we also show dominant distributions at 3% and 5% deviation using the KS-test. In Figure 8, results show the dominance of distributions for all locations from all six regions. Overall, the empirical data closely matches heavy-tailed models such as log-logistic, log-gamma, and Weibull distributions. *Surprisingly, in all cases the memoryless models such as exponential completely deviate in modeling empirical distribution.* We find that even on regions' aggregate traffic levels, the log-logistic distribution provides a good estimate of empirical data. *These results can be used as input for simulators to generate data-driven vehicular traffic traces to evaluate the performance of vehicular routing protocols.*

Next, to aid traffic visualization, we develop application to demonstrate traffic conditions on desktop and handheld

devices. In Figure 9, we show four scenarios for vehicular traffic visualization. Our aim is that to provide information about real-time traffic, future traffic conditions, locations with similar traffic, and potential traffic jam zones.

4. POTENTIAL FUTURE APPLICATION TO VEHICULAR NETWORKS

The experience gained from the analysis and modeling of traffic densities potentially aids in future design and evaluation of vehicular networks. Today, most of the simulation tools input generic or random scenarios and disregard the challenges brought by mobility in vehicular networks [2, 17, 21]. In our case, the benefit of having large library of realistic traces coupled with modeling results prove to be very helpful in developing rich scenarios for testing protocols, network dynamics, scalability of traffic, topology size estimation, and the analysis of traffic patterns. The data-driven realistic simulation tools and mobility models are necessary for accurate evaluation of vehicular routing protocols and services. However, our analysis shows that traffic characterization and communication network analysis tools(e.g. ns2) are separately developed and therefore lack a tight integration [14, 17]. Our gathering and analyzing real traffic data can aid in identifying metrics (e.g. spatio-temporal density) to develop data driven mobility models and simulators. The unique challenges (e.g. high speed, intermittent connectivity) in inter-vehicle [4] and car-to-roadside [10] communication require the development of robust and efficient routing protocols. We can use the cameras' geo-coordinates and their traffic density distribution to develop and test new performance metrics and protocols. These tests can be carried out at individual locations as well as at the city-scale level.

5. CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel framework for large-scale monitoring, analysis, and modeling of vehicular traffic using hundreds of freely available online traffic webcams. We collected imagery vehicular traffic snapshots from several different regions and converted them into traffic densities time series. We find temporal traffic patterns are stable in these regions for more than 42 days and more than 80% correlated across consecutive hours of the day. We also find empirical densities closely follow (with less than 3% and 5% deviation) heavy-tailed distributions such as log-logistic and Weibull. These results strongly indicate a revisit to use memoryless distributions to model vehicular traffic mobility. In future, we want to focus on developing realistic and data-driven models, based on these results and making this library available to the research community. Finally, we believe this activity will definitely aid in the design and evaluation of future vehicular networks and traffic engineering.

6. REFERENCES

[1] F. Bai and B. Krishnamachari. Spatio temporal variations of vehicle traffic in vanets: facts and implications. In *Vanet*, 2009.

[2] F Bai, N Sadagopan, and A Helmy. Important: A framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks. In *Infocom*, 2003.

[3] Y. Benezeth, P.M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger. Review and evaluation of commonly-implemented background subtraction algorithms. In *ICPR*, pages 1 –4, dec. 2008.

[4] J.J. Blum and et. al. Challenges of intervehicle ad hoc networks. *ITS, IEEE Tran. on*, 2004.

[5] L. Briesemeister and et. al.. In *Intelligent Vehicles Symposium*, 2000.

[6] V Bychkovsky and et. al. A Measurement Study of Vehicular Internet Access Using In Situ Wi-Fi Networks. In *Mobicom*, 2006.

[7] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. PH, 2002.

[8] A. Halati and et. al. CORSIM-corridor traffic simulation model. In *TCTS*, 1997.

[9] P. Hui and et. al. Planet scale human mobility measurement. In *ACM HotPlanet*, 2010.

[10] J Jiru and D Eilers. Car to roadside communication using ieee 802.11p technology. *Industrial Ethernet Book Issue*, 2010.

[11] N. E. Lownes and R. B. Machemehl. Vissim. In *WSC*, 2006.

[12] Q Meng and H. L. Khoo. Self-similar characteristics of vehicle arrival pattern on highways. *Jour. of Transportation Engineering*, 135(11):864–872, 2009.

[13] J. Ott and D. Kutscher. Drive-thru internet: Ieee 802.11b for "automobile" users. In *Infocom*, 2004.

[14] M. Piórkowski and et. al. Trans: realistic joint traffic and network simulator for vanets. *Sigmobile CCR*, 2008.

[15] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *PAMI*, 2005.

[16] J.P. Singh and et. al. In *VTC*, 2002.

[17] R Stanica, E Chaput, and A Beylot. Simulation of vehicular ad-hoc networks: Challenges, review of tools and recommendations. *Computer Networks*, 2011.

[18] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, 1999.

[19] N. Wisitpongphan and et. al. Routing in sparse vehicular ad hoc wireless networks. *IEEE Comm.*, 2007.

[20] J Yeo and et. al. Crawdad: a community resource for archiving wireless data at dartmouth. *Sigcomm CCR*, 2006.

[21] S Yousefi and et. al. Vehicular ad hoc networks (vanets): Challenges and perspectives. In *ITS*, 2006.

[22] X Zhang and et. al. Study of a bus-based disruption-tolerant network: mobility modeling and impact on routing. In *MobiCom*, 2007.