Statistical Models and Algorithms for Real-Time Anomaly Detection Using Multi-Modal Data

Taposh Banerjee

University of Texas at San Antonio

Joint work with
Gene Whipps (US Army Research Laboratory)
Prudhvi Gurram (US Army Research Laboratory and Booz Allen Hamilton)

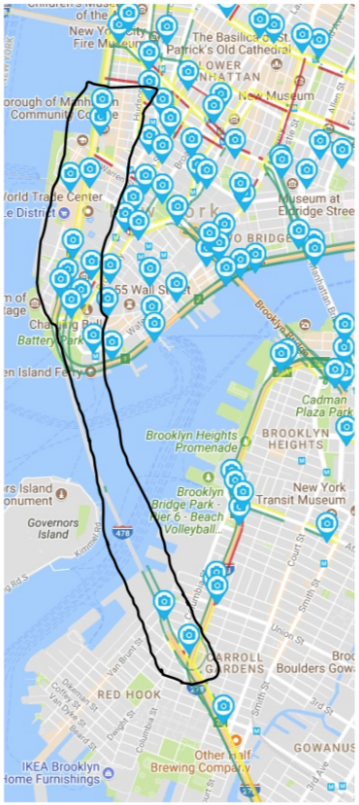
November 2, 2018

Multimodal Event Detection Problem

- Surveillance
- Infrastructure monitoring
- Environmental and natural disaster monitoring
- Crime hotspot detection for law enforcement
- Real-time traffic monitoring

- How to combine or fuse information from multiple modalities (CCTV images, Twitter, Instagram) for real-time event detection?
- What statistical models to use and algorithms to employ (optimality)?
- We take a data driven approach

Multimodal Data from a New York City Event



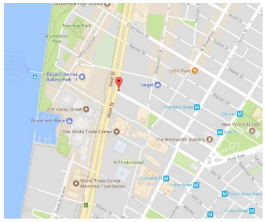
- Tunnel2Towers 5K run or walk
- CCTV data using 511ny.org
- Instagram data using picodash.com
- Twitter data using Twython

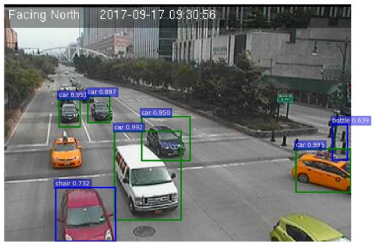
- Datasets collected around a 5K run that occurred in New York City on Sunday, September 24th, 2017
- Data also collected on two Sundays before and one Sunday after

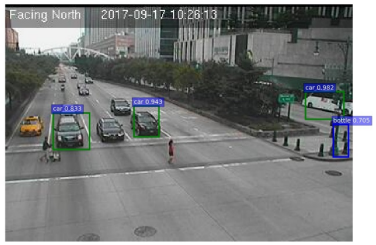
Traffic Camera Imagery from Camera on Race Path

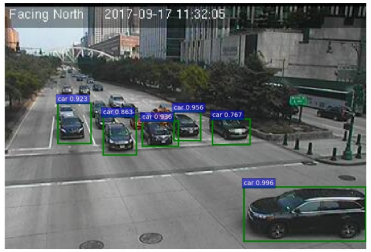
Camera: NYSDOT – 4616693 (On Race Path)

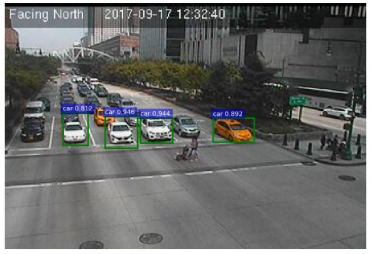
09/17/2017 (1 week before the race)

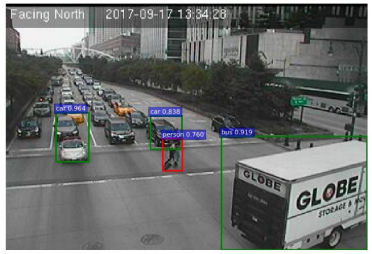










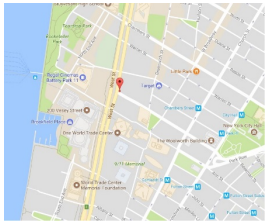


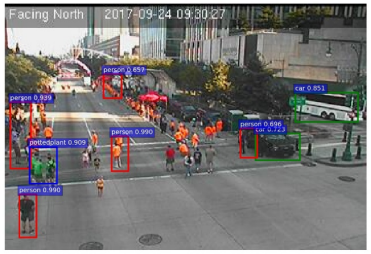
- Camera ON race path one week before

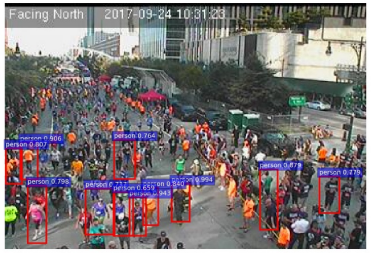
Traffic Camera Imagery from Camera on Race Path

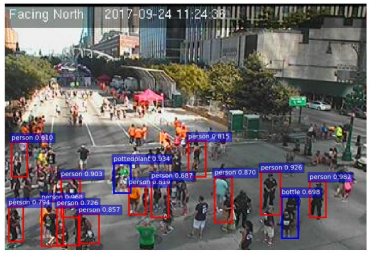
Camera: NYSDOT – 4616693 (On Race Path)

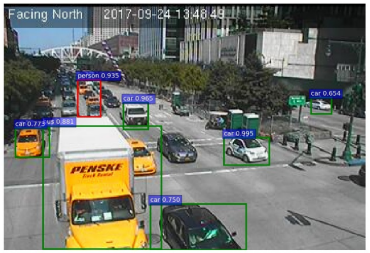
09/24/2017 (Day of the race)









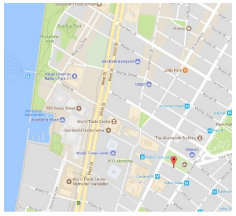


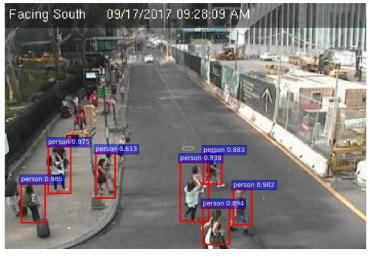
- Camera ON race path on event day
- Increase in number of persons from a week before

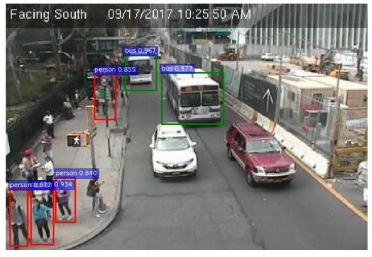
Traffic Camera Imagery from Camera off Race Path

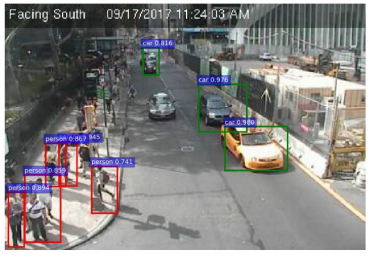
Camera: NYSDOT – 4616505 (Off Race Path)

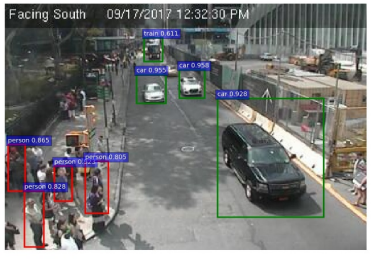
09/17/2017 (1 week before the race)

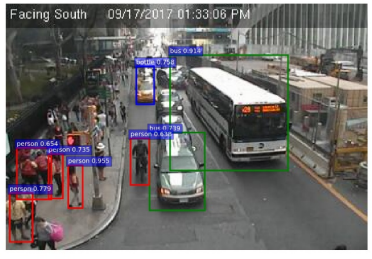










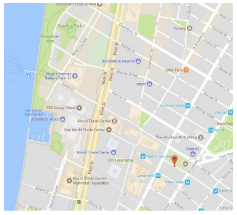


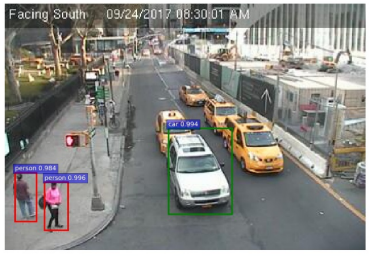
- Camera OFF race path one week before

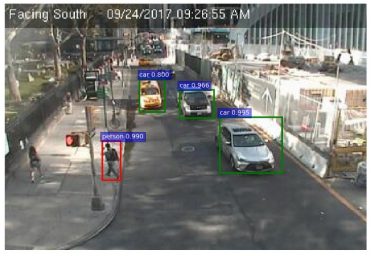
Traffic Camera Imagery from Camera off Race Path

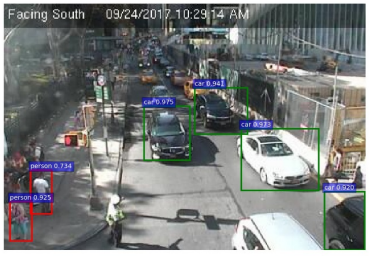
Camera: NYSDOT – 4616505 (Off Race Path)

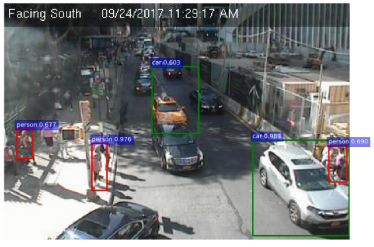
09/24/2017 (Day of the race)

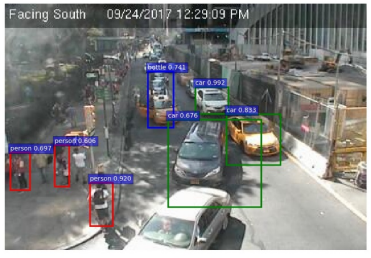










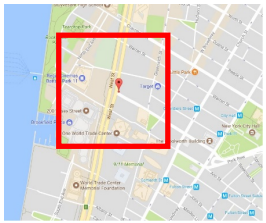


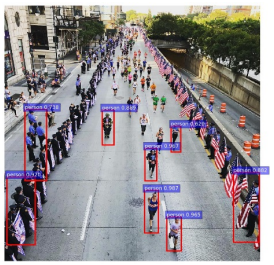
- Camera OFF race path on event day
- No significant change in number of persons

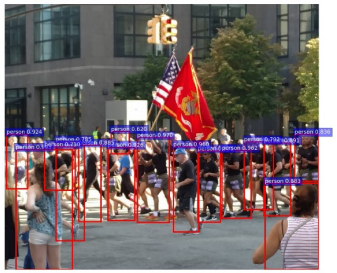
Instagram Posts from Camera on Race Path

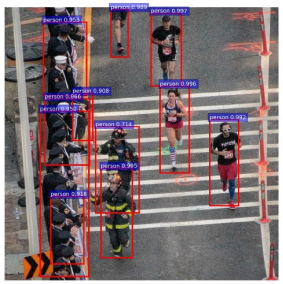
Rectangular block around NYSDOT – 4616693 (On Race Path)

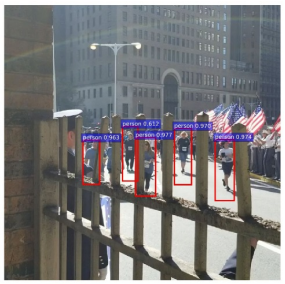
09/24/2017 (Day of the race): 295 posts in 5.5 hrs

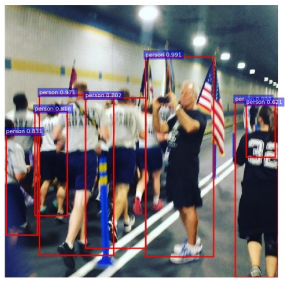


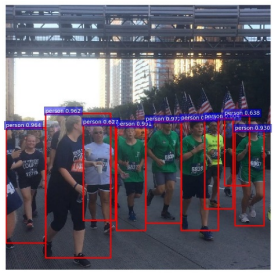










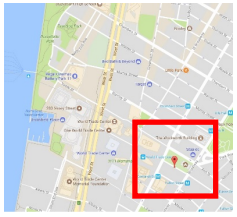


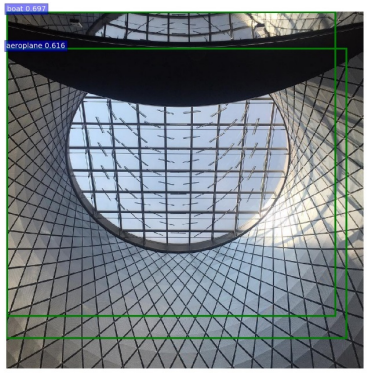
- Instagrams posts near an on path camera on event day

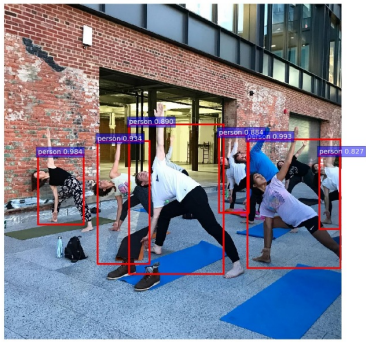
Instagram Posts from Camera off Race Path

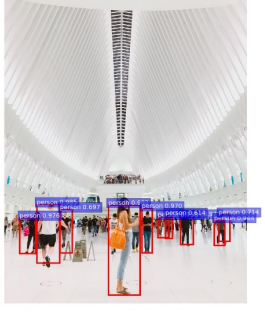
Rectangular block around NYSDOT – 4616505 (Off Race Path)

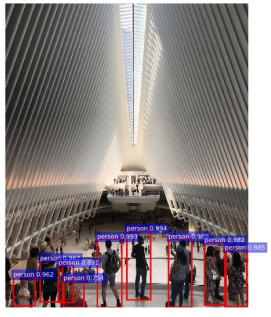
09/17/2017 (1 week before the race): 34 posts in 5.5 hrs









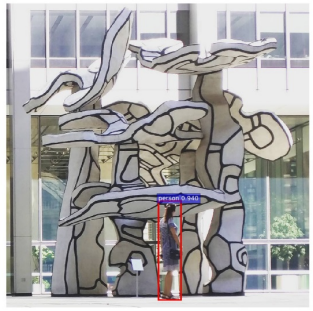


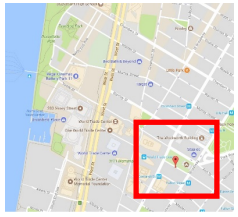
- Instagrams posts near an off path camera one week before

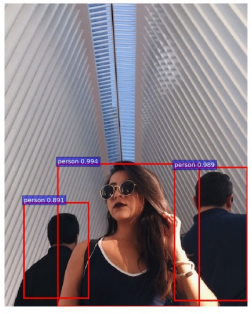
Instagram Posts from Camera off Race Path

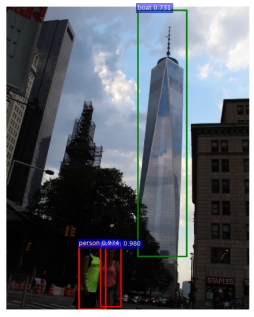
Rectangular block around NYSDOT – 4616505 (Off Race Path)

09/24/2017 (Day of the race):
38 posts in 5.5 hrs



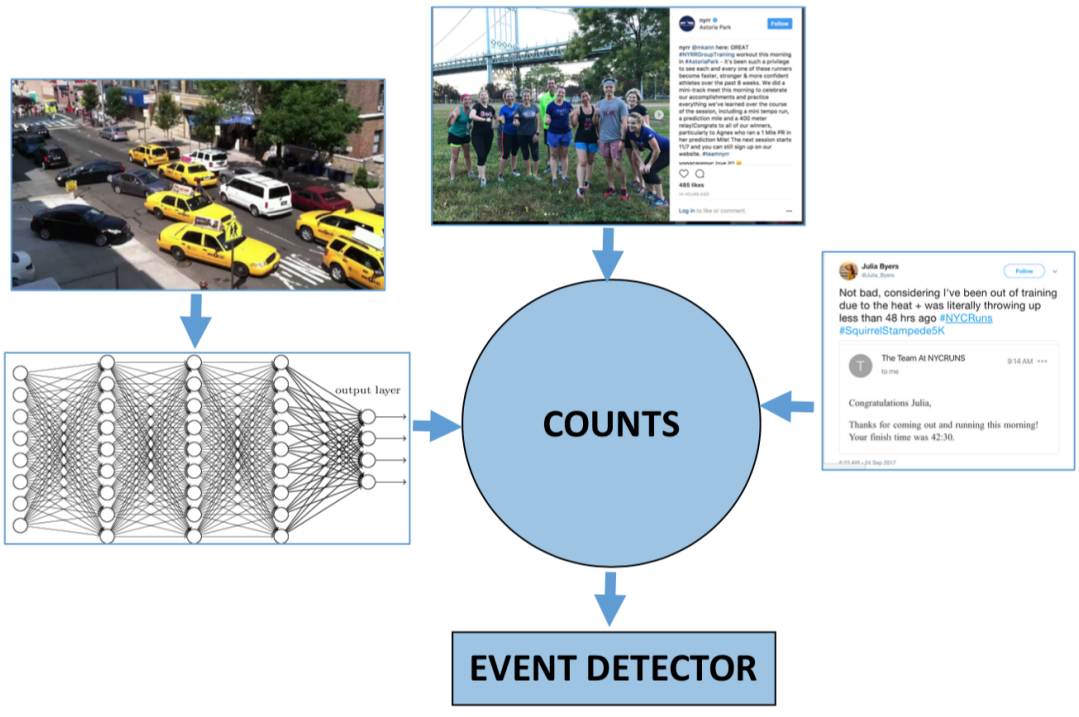




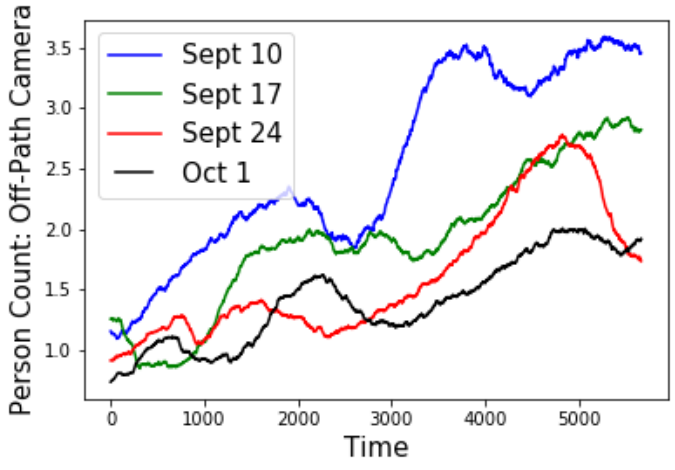


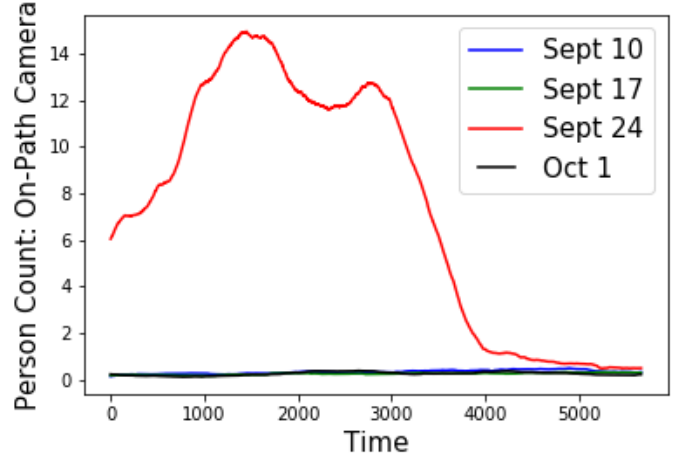
- Instagrams posts near an off path camera on event day

Extracting Counts from Multimodal Data



Average Person Counts

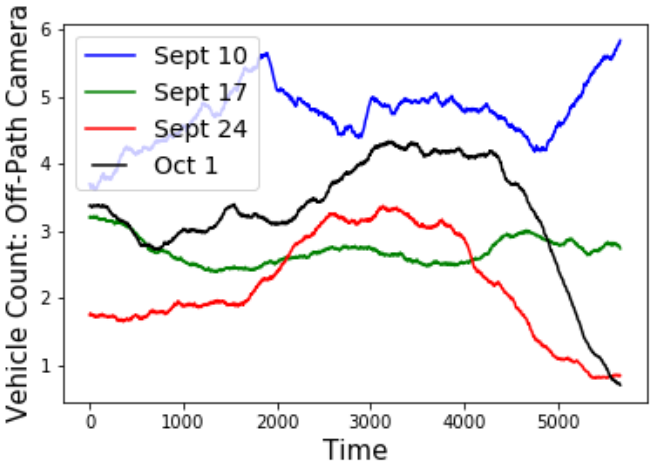


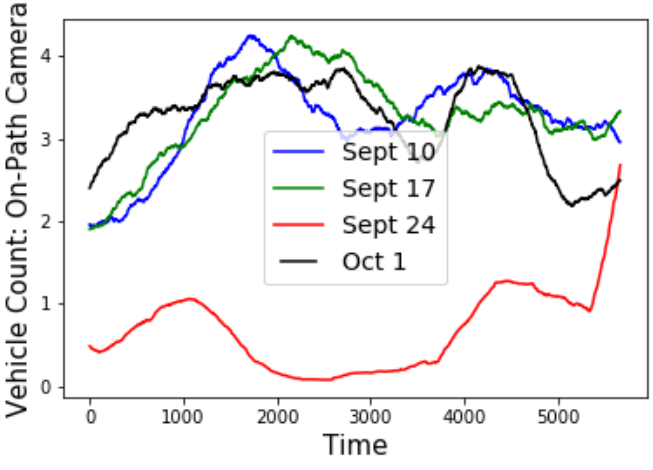


- Person counts extracted from CCTV images using convolution neural network-based object detector
- Clear increase in average count on event day for on path camera

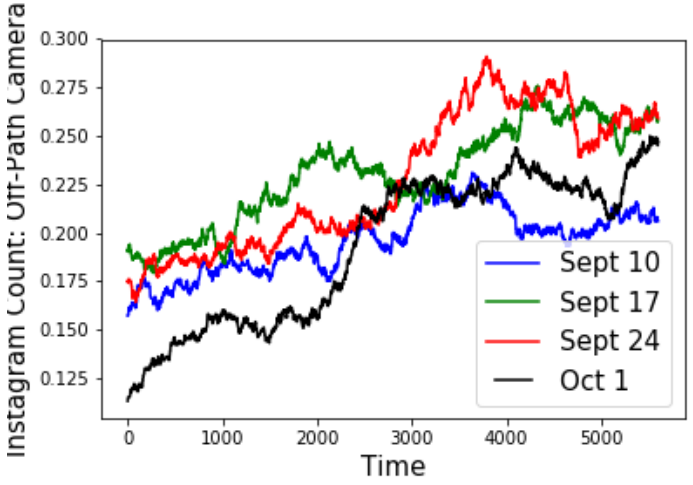
Extracting Counts from Multimodal Data

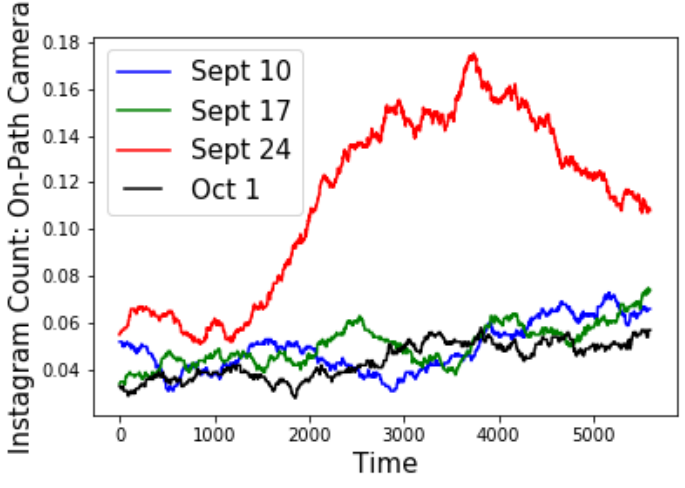
Average Vehicle Counts





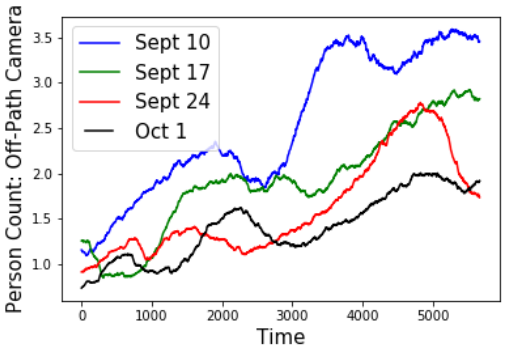
Average Instagram Counts

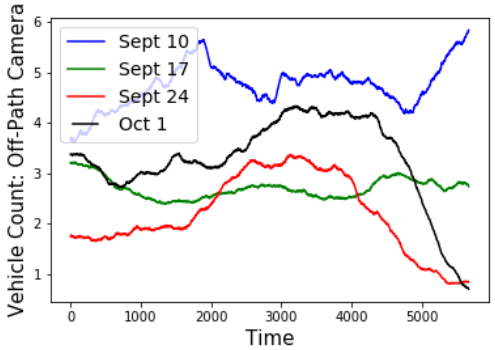


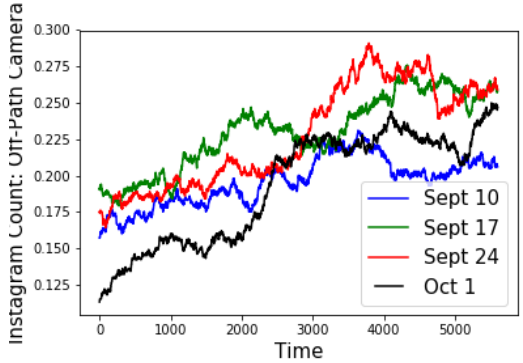


- Decrease in average vehicle counts and increase in average Instagram counts

Anomaly Detection Using Counts and Sub-Events





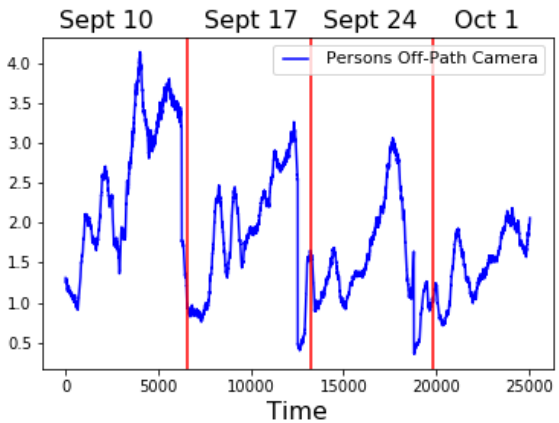


- Insights obtained using data analysis
  1. **Multiple modalities mapped to counts**: Count of sub-events can be exploited for event detection detector (also used in credit card security)
  2. **Nonstationarity**: Data is nonstationary in nature
  3. **Cyclostationarity**: Data has regular or periodic patterns
- Statistical problems
  1. **How to detect changes in levels of nonstationarity?**
  2. **How to detect changes in or deviations from learned regular behavior?**
- Sequential algorithms: detect anomaly as quickly as possible subject to constraint on false alarm rate

Periodic Statistical Behavior

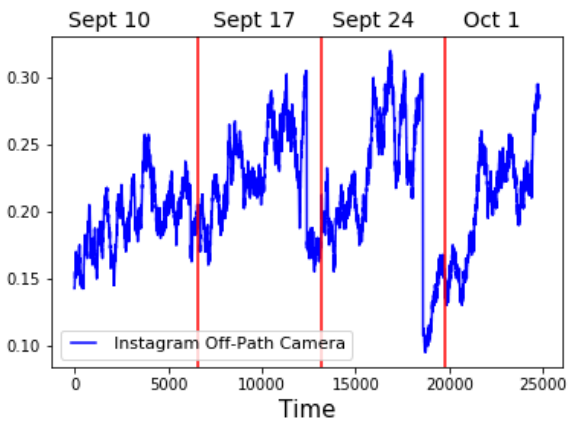
- **Average counts on four Sundays**

Person Count



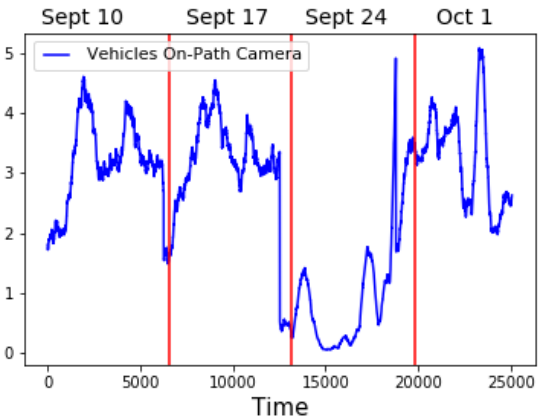
Off-path camera

Instagram Count



Off-path camera

Vehicle Count



On-path camera

- **The average counts show periodic statistical behavior**

Statistical Models

- A Bayesian model for change in parameter family
  1. A partially observable Markov decision process (POMDP) model
  2. Observations are parametrized: parameter could be mean of observations
  3. Parameters divided into two classes: normal and boundary classes
  4. Problem is to detect time at which the hidden Markov model jumps from normal to boundary states
- A non-Bayesian model for change in periodic behavior of data
  1. Define a new process: independent and periodically identically distributed (i.p.i.d.)
  2. Observations are modeled as an i.p.i.d process.
  3. Detect deviations from normal learned i.p.i.d. behavior

POMDP Model

- **States** $\{X_k\}$ a finite state Markov chain

$$X_k \in \{\mathcal{A}, 0, 1, 2, \cdots, N, N+1\}$$

  - States $\{1, 2, \cdots, N\}$ are normal and $\{0, N+1\}$ are abnormal
  - State m corresponds to mean rate λ_m with

$$\lambda_0 < \lambda_1 < \cdots < \lambda_N < \lambda_{N+1}$$

  - Detect when the Markov chain moves from normal to abnormal states
- **Controls** $\{U_k\}$ are chosen to implement the detection algorithm

$$U_k \in \{1 \text{ (stop)}, 2 \text{ (continue)}\}$$

- **Observations** Collected as long as $U_k = 2$ (continue)

$$(Y_k \mid X_k = m, U_k = 2) \sim \text{Pois}(\lambda_m), \quad m \in \{0, 1, 2, \cdots, N, N+1\}$$

POMDP Model: Continued

- **Initial distribution** π_0 for $\{X_k\}$

$$\pi_0 = (\pi_0(\mathcal{A}), \pi_0(0), \pi_0(1), \cdots, \pi_0(N), \pi_0(N+1))^T$$

which satisfies $\pi_0(\mathcal{A}) = \pi_0(0) = \pi_0(N+1) = 0$

- **Transition Probabilities**: $P_{k+1|k}(u_k) = P(u_k)$ are

$$P(2) = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & a_1 & 0 & \dots & 0 & 1-a_1 \\ 0 & p_{10} & p_{11} & \cdots & p_{1N} & p_{1(N+1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & p_{N0} & p_{N1} & \cdots & p_{NN} & p_{N(N+1)} \\ 0 & 1-a_{N+1} & 0 & \dots & 0 & a_{N+1} \end{bmatrix}$$

$$P(1) = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{bmatrix}$$

POMDP Model: Continued

- **Cost** $C(x,u)$ associated with state $X = x$ and control $U = u$

$$C(x,1) = C_f^T e_x = (0,0,c_f,c_f,\cdots,c_f,0)\ e_x \quad \text{[FALSE ALARM]}$$
$$C(x,2) = C_d^T e_x = (0,c_d,0,0,\cdots,0,c_d)\ e_x \quad \text{[DELAY]}$$

- **Cost to go** for policy $\Phi = \{u_k = \phi_k(I_k)\}$

$$V(\pi_0) = \min_{\Phi} \mathbb{E}\left[\sum_{k=1}^{\infty} C(x_k,u_k)\right]$$

- **Bellman's equation** satisfied by value function $V(\pi)$

$$V(\pi) = \min\left\{C_f^T\pi,\ C_d^T\pi\ +\ \sum_y V(T(\pi,y,2))\ \sigma(\pi,y,2)\right\}$$

Optimal Policy for POMDP

- Optimal policy a function of belief state $\pi_k = \mathbb{P}(X_k = x_k | I_k)$
- π_k can be computed recursively using emission probabilities $B_y(u)$

$$\pi_{k+1} = T(\pi_k, y_{k+1}, u_k) = \frac{B_{y_{k+1}}(u_k)\ P(u_k)^T \pi_k}{\mathbf{1}^T B_{y_{k+1}}(u_k)\ P(u_k)^T \pi_k}$$

- Convexity of stopping region

Theorem

Optimal policy is stationary and stopping region

$$R_1 = \{\pi : \mu^*(\pi) = 1\}$$

is convex

- Standard conditions to establish threshold structure NOT satisfied
  - Total positivity conditions for transition structure and emission probabilities: *satisfied*
  - Monotonicity and submodularity structure of cost function: *not satisfied*

Special Structure of POMDP Optimal Policy

Theorem (Post-change absorbing transitions)

Let

$$\bar{P} = \begin{bmatrix} p_{10} & p_{11} & \cdots & p_{1N} & p_{1(N+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{N0} & p_{N1} & \cdots & p_{NN} & p_{N(N+1)} \end{bmatrix}$$

If rows of $\bar{P}$ are identical and $a_1 = a_{N+1} = 1$, then optimal policy depends only on components $\pi(0)$ and $\pi(N+1)$ of the belief state π

Theorem (Post-change random transitions)

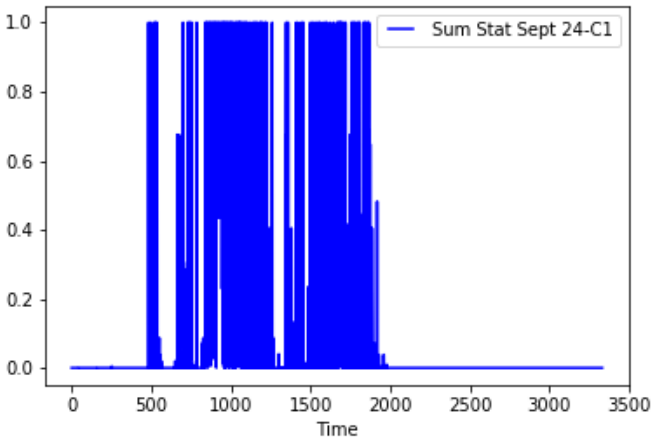
If rows of $\bar{P}$ are identical and $a_1 = a_{N+1} = 1/2$, then optimal policy depends only on $\pi(0) + \pi(N+1)$ of the belief state π

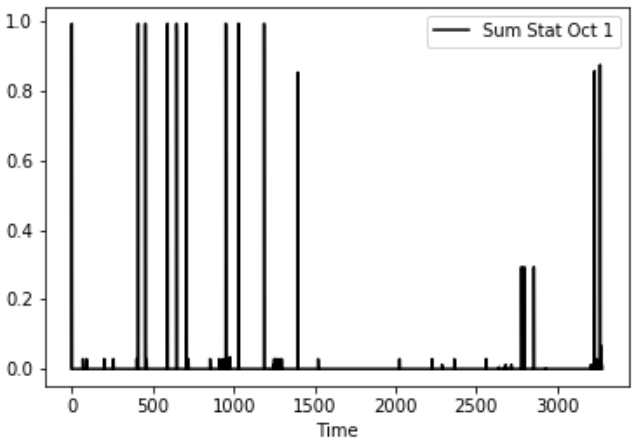
- With condition $a_1 = a_{N+1} = 1/2$ problem reduces to classical quickest change detection test of Shiryaev and Kolmogorov

Application to Real Data: Belief Sum Algorithms

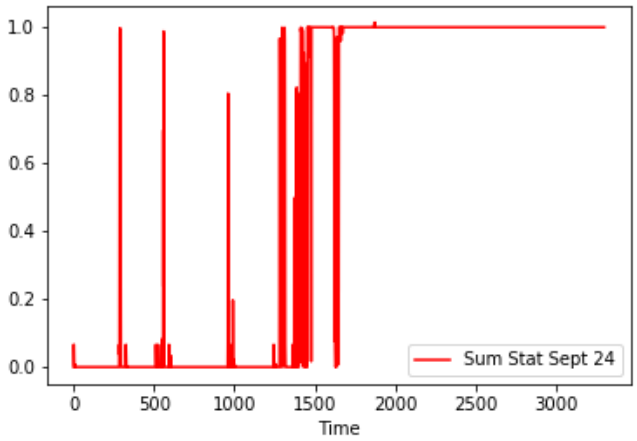
Baseline learned from the first day of data using Poisson modeling

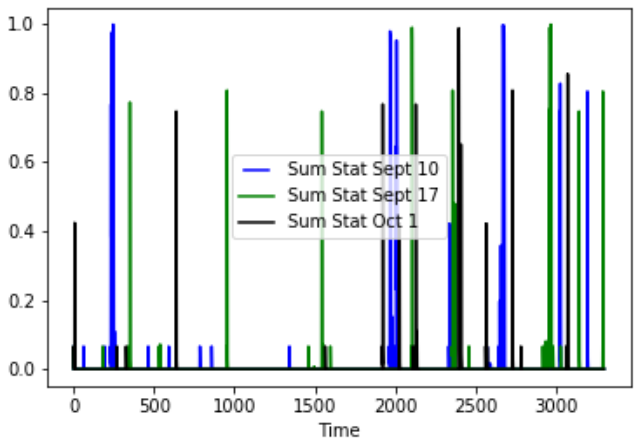
Person Counts





Instagram





A Model to Capture Periodic Statistical Behavior

- Recall that an independent and identically distributed (i.i.d.) process is a sequence of random variables that are independent and have the same distribution.
- We define a new category of stochastic processes called **independent and periodically identically distributed (i.p.i.d.)** processes:

Definition

Let $\{X_n\}$ be a sequence of random variables such that the variable X_n has density f_n. The stochastic process $\{X_n\}$ is called independent and periodically identically distributed (i.p.i.d) if X_n are independent and there is a positive integer T such that the sequence of densities $\{f_n\}$ is periodic with period T:

$$f_{n+T} = f_n, \quad \forall n \geq 1.$$

We say that the process is i.p.i.d. with the law $(f_1, \cdots, f_T)$.

- An i.p.i.d. process is cyclostationary. T can be interpreted as the number of samples in a day or week

Change Point Model

- Consider another periodic sequence of densities $\{g_n\}$ such that

$$g_{n+T} = g_n, \quad \forall n \geq 1.$$

- These densities need not be all different from the set of densities $(f_1, \cdots, f_T)$, but we assume that there exists at least an i such that they are different:

$$g_i \neq f_i, \quad \text{for some } i = 1, 2, \cdots, T.$$

- **Our change point model is**: there exists $\nu \in \mathbb{N}$ such that

$$X_n \sim \begin{cases} f_n, & \forall n < \nu, \\ g_n & \forall n \geq \nu. \end{cases}$$

- This model is equivalent to saying that we have two i.p.i.d. processes, one governed by the densities $(f_1, \cdots, f_T)$ and another governed by the densities $(g_1, \cdots, g_T)$, and at the change point ν, the process switches from one i.p.i.d. process to another

An Algorithm For Detecting Changes in i.p.i.d. models

- We define the following algorithm and called the **Periodic-CUSUM** algorithm: compute the sequence of statistics

$$W_{n+1} = \max_{1 \le k \le n+1} \sum_{i=k}^{n+1} \log \frac{g_i(X_i)}{f_i(X_i)}$$

and raise an alarm as soon as the statistic is above a threshold A:

$$\tau_c = \inf\{n \ge 1 : W_n > A\}$$

- **Why this statistic?** $\sum_{i=k}^{n+1} \log \frac{g_i(X_i)}{f_i(X_i)}$ is the logarithm of the likelihood ratio between observations $X_1, \cdots, X_{n+1}$, given that the change point $\nu = k$. Since, we do not know if $\nu = k$, we use its maximum likelihood estimate.
- We call it periodic-CUSUM because for $T = 1$ this algorithm reduces to a famous CUSUM algorithm in the literature

Recursive Computation of Periodic-CUSUM Statistic

Theorem

The statistic sequence $\{W_n\}$ *can be recursively computed as*

$$W_{n+1} = W_n^+ + \log \frac{g_{n+1}(X_{n+1})}{f_{n+1}(X_{n+1})},$$

where $(x)^+ = \max\{x, 0\}$. *Further, since the set of pre- and post-change densities* $(f_1, \cdots, f_T)$ *and* $(g_1, \cdots, g_T)$ *are finite, the recursion can be computed using finite memory needed to store these* $2T$ *densities, past statistic, and current observation.*

- **In general, in change point literature, it is rare to find recursively computable algorithms that are also optimal in a well-defined sense. Is Periodic-CUSUM optimal in any sense?**

Stochastic Optimization Problem For Change Detection

- **Change point**: the change point is ν and unknown
- **Stopping time**: a positive integer valued random variable τ is called a stopping time if the decision to stop at time n $(\tau = n)$ is only a function of observations until time n, $(X_1, \cdots, X_n)$
- **Pollak's (1985) formulation**:

$$\begin{aligned} \min_{\tau} \quad & \sup_{\nu \geq 1} \mathsf{E}_{\nu}[\tau - \nu | \tau \geq \nu] \\ \text{subj. to} \quad & \mathsf{E}_{\infty}[\tau] \geq \beta, \end{aligned}$$

- **Lorden's (1971) formulation**

$$\begin{aligned} \min_{\tau} \quad & \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathsf{E}_{\nu}[\tau - \nu | X_1, \cdots, X_{\nu-1}] \\ \text{subj. to} \quad & \mathsf{E}_{\infty}[\tau] \geq \beta, \end{aligned}$$

- **These are the two most famous optimization problems in the literature**

Lower Bound On Any Stopping Time

- Let $I = \frac{1}{T}\sum_{i=1}^{T} D(g_i \parallel f_i)$ where

$$D(g_i \| f_i) = \int_x g_i(x) \log \frac{g_i(x)}{f_i(x)} dx.$$

is the Kullback-Leibler divergence between the densities g_i and f_i.

Theorem

If $0 < I < \infty$, then for any stopping time τ satisfying the false alarm constraint $\mathsf{E}_\infty[\tau] \geq \beta$, we have as $\beta \to \infty$

$$\sup_{\nu \geq 1} \operatorname{ess\,sup} \mathsf{E}_\nu[\tau - \nu | X_1, \cdots, X_{\nu-1}]$$

$$\geq \sup_{\nu \geq 1} \mathsf{E}_\nu[\tau - \nu | \tau \geq \nu] \geq \frac{\log \beta}{I}(1 + o(1)),$$

where an $o(1)$ term is one that goes to zero in the limit as $\beta \to \infty$.

Optimality of Periodic-CUSUM

Theorem

Let the information number I satisfy $0 < I < \infty$. Then, the Periodic-CUSUM stopping time τ_c with $A = \log \beta$ satisfies the false alarm constraint

$$\mathsf{E}_\infty[\tau_c] \geq \beta,$$

and as $\beta \to \infty$,

$$\begin{aligned}
&\sup_{\nu \geq 1} \mathsf{E}_\nu[\tau_c - \nu | \tau_c \geq \nu] \\
&\quad \leq \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathsf{E}_\nu[\tau_c - \nu | X_1, \cdots, X_{\nu-1}] \\
&\quad \leq \frac{A}{I}(1 + o(1)) = \frac{\log \beta}{I}(1 + o(1)).
\end{aligned}$$

Unknown Post-Change Model

- **What if the post-change i.p.i.d. distributions are unknown?**
- Assume that the post-change i.p.i.d. law belongs to a **finite** set of M possible distributions

$$(g_1^{(1)}, \cdots, g_T^{(1)}), \cdots, (g_1^{(M)}, \cdots, g_T^{(M)})$$

- Compute a Periodic-CUSUM statistic for each possible post-change hypothesis:

$$W_{n+1}^{(\ell)} = \left(W_n^{(\ell)}\right)^+ + \log \frac{g_{n+1}^{(\ell)}(X_{n+1})}{f_{n+1}(X_{n+1})}$$

$$\tau_{c\ell} = \inf\left\{n \geq 1 : W_n^{(\ell)} \geq \log(\beta M)\right\}$$

- Use the following first-stopping rule

$$\tau_{cm} = \inf\left\{n \geq 1 : \max_{1 \leq \ell \leq M} W_n^{(\ell)} \geq \log(\beta M)\right\} = \min_{\ell} \tau_{c\ell}.$$

which is the first time any of the ℓ Periodic-CUSUMs raise an alarm

Optimality of first-stopping rule

Theorem

The false alarm constraint is satisfied:

$$\mathsf{E}_\infty[\tau_{cm}] \geq \beta.$$

Further, if $(g_1^{(\ell)}, \cdots, g_T^{(\ell)})$ *is the true post-change i.p.i.d. law and* $I_\ell = \frac{1}{T}\sum_{i=1}^{T} D(g_i^{(\ell)} \parallel f_i)$, *then*

$$\begin{aligned}\sup_{\nu \geq 1} \mathsf{E}_\nu[\tau_{cm} - \nu | \tau_{cm} \geq \nu] &\leq \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathsf{E}_\nu[\tau_{cm} - \nu | X_1, \cdots, X_{\nu-1}] \\ &\leq \frac{\log \beta}{I_\ell}(1 + o(1)).\end{aligned}$$

- The stopping rule τ_{cm} is thus asymptotically optimal with respect to the criteria of Lorden and Pollak, **uniformly** over each possible post-change hypothesis $(g_1^{(\ell)}, \cdots, g_T^{(\ell)})$, $\ell = 1, \cdots, M$.

Parametric i.p.i.d. Models

- Learning an i.p.i.d. model mean learning the T densities $(f_1, \cdots, f_T)$ and possibly $(g_1, \cdots, g_T)$.
- **Learning entire distributions is hard**. This motivates the following definition

Definition

A stochastic process $\{X_n\}$ is called a parametric i.p.i.d. process if

$$X_n \overset{\text{ind}}{\sim} p(\cdot; \theta_n), \ \forall n$$
$$\theta_n = \theta_{n+T}, \ \forall n.$$

- Learning an i.p.i.d. model is then equivalent to learning a finite set of T parameters $(\theta_1, \cdots, \theta_T)$.

Step Model for Parameters

- parametric i.p.i.d. model: too many parameters to learn? **If sampling frequency is high and T corresponds to a week, T may be in thousands**

$$X_n \overset{\text{ind}}{\sim} p(\cdot;\theta_n),\ \forall n$$
$$\theta_n = \theta_{n+T},\ \forall n.$$

- Divide parameters $\{\theta_k\}_{k=1}^{T}$ into batches

$$\overbrace{\theta_1,\cdots,\theta_{N_1}}^{\theta_{B_1}},\overbrace{\theta_{N_1+1},\cdots,\theta_{N_2}}^{\theta_{B_2}},\cdots,\overbrace{\theta_{N_{E-1}+1},\cdots,\theta_{N_E}}^{\theta_{B_E}}$$

- **Assume step model**: parameters constant within a batch (e.g. an hour)

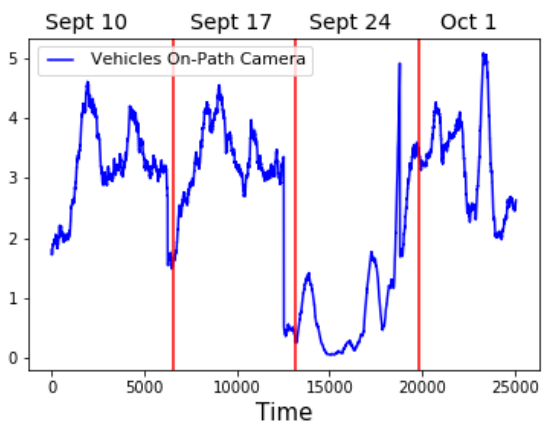
$$\overbrace{\theta^{(1)},\cdots,\theta^{(1)}}^{\theta_{B_1}},\overbrace{\theta^{(2)},\cdots,\theta^{(2)}}^{\theta_{B_2}},\cdots,\overbrace{\theta^{(E)},\cdots,\theta^{(E)}}^{\theta_{B_E}}$$

- Learn parameters $\theta^{(1)}$ to $\theta^{(E)}$ from data and detect deviations from it in real-time: $E \ll T$

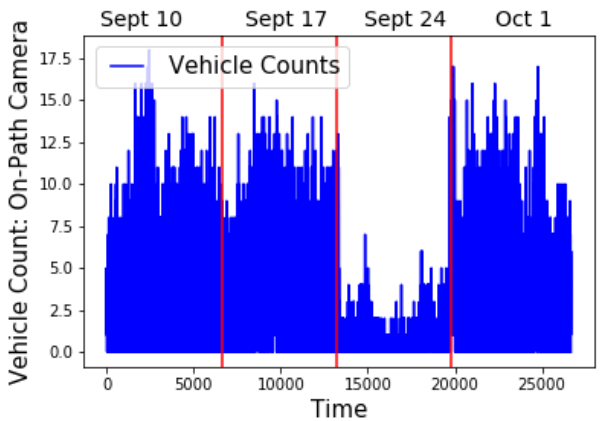
Application to Vehicle Data

- **Baseline learned from first day of data using Poisson modeling. Post-change parameter is set to half of baseline.**

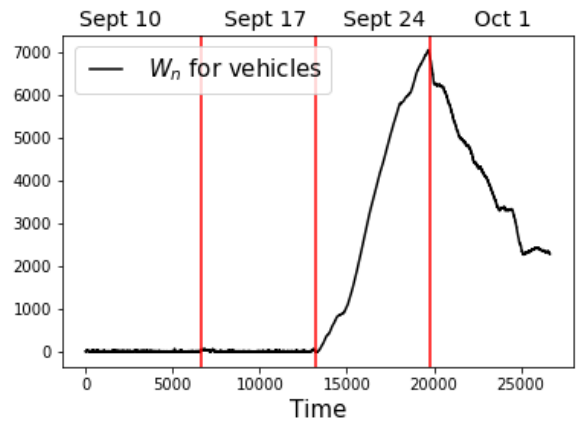
Average Vehicle Counts



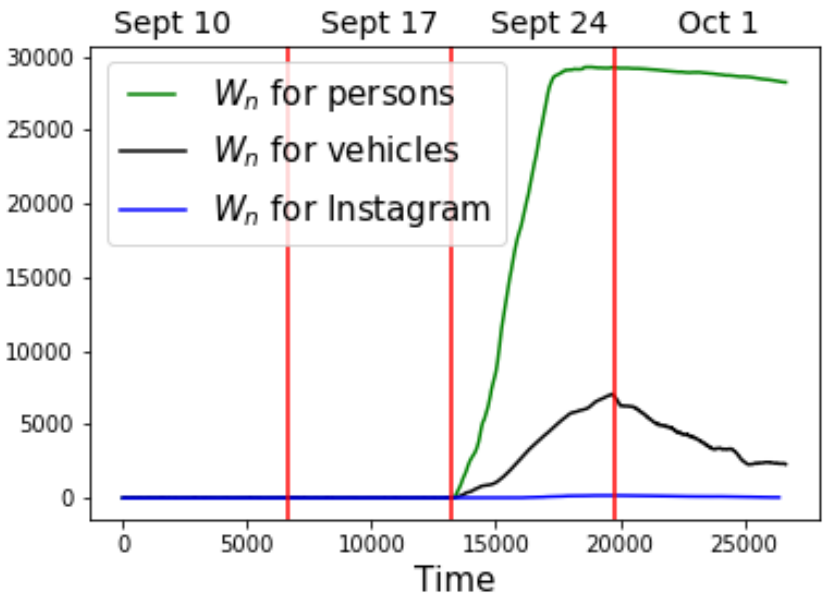
Vehicle Counts

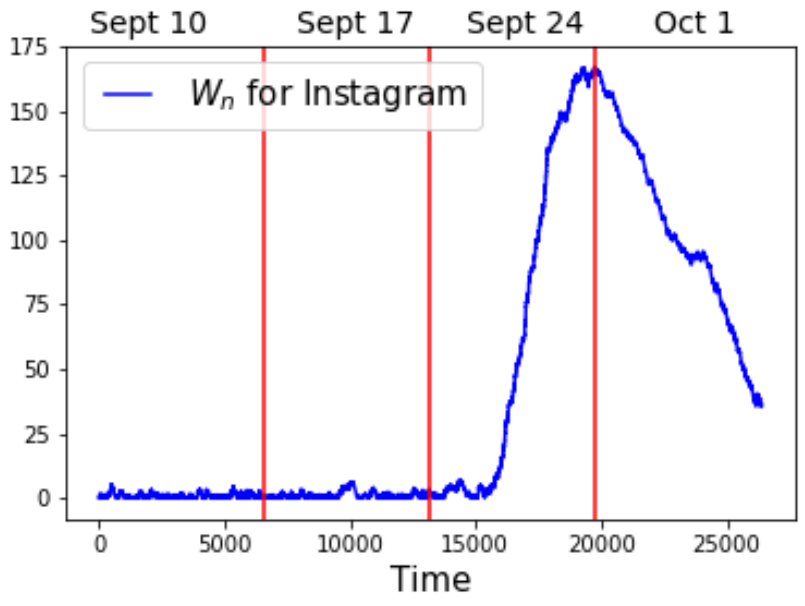


Periodic-CUSUM Statistic



Application to Real Data: Periodic-CUSUM Algorithm





Conclusions

- **New Framework for Multimodal Signal Processing**: Extracted counts of objects or sub-events from the data to convert multimodal data into a single modality
- **New Statistical Models for Detection**: Developed new POMDP formulation and defined new stochastic process family (i.p.i.d) to capture nonstationary processes
- **New Algorithms and Optimality**: Obtained algorithms that are optimal with respect to well-defined criteria
- **Application to Real Data**: Applied the developed algorithms to data collected from NYC after learning the baseline from the first day of data collection

Publications

The results in this talk can be found in the following papers:

1. T. Banerjee, P. Gurram, and G. Whipps, "**Quickest Detection Of Deviations From Periodic Statistical Behavior**, " submitted to ICASSP 2019. https://arxiv.org/abs/1810.12760.
2. T. Banerjee, G. Whipps, P. Gurram, and V. Tarokh, "**Cyclostationary Statistical Models and Algorithms for Anomaly Detection Using Multi-Modal Data**," In IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2018. https://arxiv.org/abs/1807.06945
3. T. Banerjee, G. Whipps, P. Gurram, and V. Tarokh, "**Sequential event detection using multimodal data in nonstationary environments**," In Proc. of International Conference on Information Fusion (FUSION), Cambridge, UK, July, 2018. https://arxiv.org/abs/1803.08947

References

1. V. Krishnamurthy, Partially Observed Markov Decision Processes. Cambridge University Press, 2016.
2. D. P. Bertsekas and S. Shreve, Stochastic optimal control: the discretetime case. Academic Press, 1978.
3. V. Krishnamurthy, Bayesian sequential detection with phase-distributed change time and nonlinear penaltya pomdp lattice programming approach, IEEE Transactions on Information Theory, vol. 57, no. 10, pp. 70967124, 2011.
4. V. V. Veeravalli and T. Banerjee, Quickest Change Detection. Academic Press Library in Signal Processing: Volume 3 Array and Statistical Signal Processing, 2014.
5. H. V. Poor and O. Hadjiliadis, Quickest detection. Cambridge University Press, 2009.
6. A. G. Tartakovsky, I. V. Nikiforov, and M. Basseville, Sequential Analysis: Hypothesis Testing and Change-Point Detection. Statistics, CRC Press, 2014.
7. W. S. Lovejoy, On the convexity of policy regions in partially observed systems, Operations Research, vol. 35, no. 4, pp. 619621, 1987.
8. A. N. Shiryayev, Optimal Stopping Rules. New York: Springer-Verlag, 1978.

The End