Emulating Smart City Sensors Using Soft Sensing and Machine Intelligence: a Case Study in Public Transportation

Cem Kaptan, *Student Member, IEEE*, Burak Kantarci, *Senior Member, IEEE* Tolga Soyata, *Senior Member, IEEE*, Azzedine Boukerche, *Fellow, IEEE*,

***Abstract*—This paper proposes a new framework for emulating the functionality of a sensor by using multiple available soft sensors and machine intelligence algorithms. As a case study, the localization of city buses in a smart city setting is investigated by using the accelerometer and microphones of the passengers and a Support Vector Machine (SVM) running in the cloud; in this application, the GPS functionality is emulated by using these two soft sensors. What makes such an emulation feasible is the statistical dependence of the location data (which would normally be obtained from a GPS) on the accelerometer and microphone data; while accelerometers capture data that relate to the typical stop/start patterns of the buses, microphone capture enter/exit patterns of the passengers through the sound levels inside the bus. We evaluate our proposed scheme through simulations and show that the proposed framework can operate with more than 90% accuracy in estimating the location of public buses while preserving the actual location privacy of the smartphone users. This approach results in smartphone battery energy savings of 38–46% (as compared to GPS-based approaches) due to the elimination of the power-hungry GPS devices.**

***Index Terms*—5G; analytics, crowd-sensing, dedicated sensors; non-dedicated sensors; smart cities, smart transportation**

I. Introduction

As we move towards the 5G era, huge volume of data exchange at significantly low latency and at higher reliability paves the way for wide adoption of the Internet of Vehicles (IoV) concept [1]. Cooperative transmission with the help of cooperating small cell Base Stations makes it possible to improve the transmission capacity as well as reliability in 5G networks [2]. In the IoV concept, cooperation among the vehicles is of paramount importance. With the advent of 5G mobile cellular networks, scalability and latency challenges experienced by IEEE 802.11p and Long-Term Evolution (LTE)-enabled vehicular networks will be addressed, particularly for real time multimedia sharing [3]. Evolving 5G communication models have enabled offloading and opportunistic spectrum sharing for vehicular nodes to provide real time infotainment services more efficiently [4]. As the connected vehicles have become a reality, vehicular sensing has appeared as a viable solution to assist in several services —in addition to infotainment services— in smart cities such as traffic management, forensics, and air pollution control [5].

Wide usage of smartphones and the CAN-Bus interfaces in vehicles can be consolidated into a semi-dedicated sensing platform so that the smartphones can read the data through the interface and transform the data into a usable format by vehicular services [6]. However, localization and tracking of vehicles is a crucial problem in a smart environments [7]. For instance, under hostile settings where GPS signal quality is not of high quality, a smart localization approach has to be used by connected vehicles. This localization method can be either learning-based [8], or it can be a hybrid solution of hardware and software such as utilizing passive RFID tags in the vehicles and collaborative cognition and correction of localization errors [9]. Furthermore, in order to improve the localization accuracy and improve the performance of Global Navigation Satellite Systems, sensor fusion techniques such as cubature Kalman Filter can also be utilized over the collaboratively observed data transmitted via DSRC links [10]. The state of the art mainly focuses on replacing GPS, however public transportation vehicles almost always move over static routes. Therefore, tracking public transport vehicles requires localization of that vehicle in a cloaked (stational) region rather than a precise GPS localization. Furthermore, in a smart city setting, using citizens as sensors [11], [12] or service providers [13] can also reduce capital and operational expenditures. Sensor emulation and coverage has been studied in the literature by a few researchers [14], [15]. Application areas in sensor emulation can be various. As an application area we have identified smart transportation.

Based on the motivations stated above, in this paper, we propose a participatory GPS-less solution for the localization of public transportation vehicles moving over static routes. Our proposed framework relies on two primary building blocks: 1) Multi-dimensional crowd-sensed data that is acquired from the built-in sensors of the smartphones and/or tablets carried by the passengers, 2) Online prediction engine that couples probabilistic location regions using a machine learning-based classification. Through simulations, we show that the location of a bus can be estimated by the proposed framework with a 90% accuracy. Furthermore, we also show that instead of aggregating participatory sensed GPS data, which would have been possible, acquiring ambient and kinematic sensor data can lead to some 38–46% savings in the battery energy of participating devices. Besides the benefits of the proposed

C. Kaptan, B. Kantarci and A. Boukerche are with the School of Electrical Engineering and Computer Science at the University of Ottawa, Ottawa, ON, K1N 6N5, Canada. E-mail: {ckapt046, burak.kantarci, aboukerc}@uottawa.ca.

T. Soyata is with the Department of Electrical and Computer Engineering, SUNY Albany, Albany NY, 12203 E-mail: tsoyata@albany.edu.

978-1-5386-3180-5/18/$31.00 ©2018 IEEE

framework in terms of prediction accuracy and energy savings, avoiding exact GPS data in the sensory acquisition stage also ensures location privacy of the participants.

The rest of the paper is organized as follows. In Section II, we present the related work and state of the art in the localization of vehicles in smart transportation. In Section III, we present our proposed framework in detail. In Section IV, we present and discuss numerical results regarding the performance of our approach. We finalize our paper and provide future directions in Section V.

II. Related Work and Motivation

A. Related Work

Localization problem has been widely studied in the context of wireless sensor networks [16]–[18]. In smart city setting, adopting WSN localization principles is beneficial; on the other hand, localization and tracking of vehicles introduces additional environmental and contextual constraints. The study in [19] evaluates the impact of internal time discrepancy and data loss on extracting passengers' origin information in a cellular network-based transportation system. The authors showed that the location data loss events have a strongly and positively correlated with shadow effects in urban areas.

A deep learning model is proposed in [20] to forecast the destinations of bus passengers from entry-only smart-card data and land-use characteristics. Training their deep learning model was achieved using a large amount of transactional data collected from the Automatic Fare Collection (AFC) system along with land-use characteristics

In [21], the authors proposed an architecture where actionable knowledge could be extracted from IoT data streams coupled with historical IoT data. The proposed architecture enables real time processing of events via statistical processing and/or machine learning algorithms depending on the context and application in smart city settings.

An online traffic prediction system has been proposed based on the acquired traffic data [22] with the objective of early congestion detection in order to actuate citizens to help reduce the likelihood and/or the duration of congestion.

The authors in [23] propose a new mechanism for the parallel management and control system (PTMS) which builds on social signal, agent control, IoT, and having ACP (i.e. artificial system, computational experiment and parallel execution) at the core of the proposed system.

Location privacy is an important concern in vehicular networks, especially in vehicular cloud systems. For instance, passengers in a vehicle could use their cellular data to upload collected records to the cloud; if they wish to save their data package they will be charged some amount depending on which service provider they are using. The study in [24] defines privacy as the amount of information revealed to the service provider. In addition to the revealed information, location privacy is a crucial challenge in smart transportation when user involvement is inevitable. To address this challenge, the authors in [25] propose a Dynamic Group Division (DGD) algorithm integrated with a 5G-based Vehicular Sensor Network architecture to remarkably improve the location and trajectory privacy of the vehicles.

Sensory data acquisition and analysis through machine learning techniques has been studied in [26] to determine the transportation mode of a person. Sensors that were used in data acquisition were accelerometers, gyroscopes, rotation vectors, and GPSs. Through a comparative study on tree-based models (e.g. single decision tree, bagging, or random forest) and SVM, GPS was aimed to be eliminated when determining the transportation mode due to its lack of availability in rural and closed areas, and high energy consumption.

B. Motivation

While the last reviewed study can be considered as the closest one to ours, there are significant differences between our proposed approach and the research in [26]. The study in [26] aims to obtain the transportation mode of a person, along with possible elimination of GPS data; alternatively, our proposed solution targets to obtain crowd-sensed data from multiple built-in sensors to estimate the location of a public transportation vehicle within the bounds of a region. Moreover, as we focus on vehicles moving on static routes, while eliminating GPS usage, by estimating a bounded region as the location of the vehicle, we protect location privacy of the participants who provide their sensor readings.

III. Proposed Solution

Our proposed framework is illustrated in Fig. 1, which consists of four main subsystems: **1)** Data acquisition, **2)** Data aggregation, **3)** Probabilistic prediction, and **4)** Analytics. Acquired data is stored on a remote server, on which analytics and probabilistic prediction algorithms run. Table I presents the notation used in the description of the proposed framework.

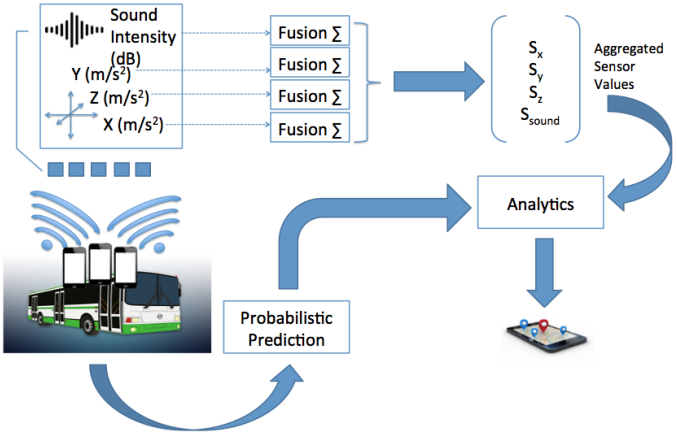

Fig. 1: GPS-less localization architecture. Proposed architecture consists of four sub-systems: **1)** Data acquisition, **2)** Data aggregation, **3)** Probabilistic prediction, and **4)** Analytics. A remote server is dedicated to the storage of acquired sensory data, which can also be used to run the analytics and probabilistic prediction procedures.

Below is a brief explanation of these subsystems:

1) Data acquisition: mainly involves the recruitment of passengers for participatory sensing of ambient and kinematic data. Here, a possible challenge might be the quantification of the trustworthiness of the acquired data [27]. However, by making a fundamental assumption, we model each participant/passenger with rational and cooperative behavior which results in acquired data that does not interfere with the value of the aggregated data beyond possible noise factors.

TABLE I: Notation used in the paper. The variables and constants are grouped in three categories: The first category (top section of the table) denotes the objects, i.e., sensors, sensory detection zones and coordinates, second category includes the variables regarding aggregated sensed values, the third category is related to the variables regarding communications, and the fourth category (bottom section of the table) is related to participant recruitment scenarios.

Parameter	**Value**
k	Index of a window of ℓ landmarks
v	Index of a vehicle
τ	Time (discrete) of interest
r_τ^v	The sensory detection zone (SDZ) of vehicle v at time τ
z_i	SDZ i
z_{i_x}	x coordinate of the SDZ i
z_{i_y}	y coordinate of the SDZ i
z_{i_z}	z coordinate of the SDZ i
ℓ	# of possible landmarks that a vehicle can be present
$\mathfrak{T}_k^v$	k^{th} sliding window for set of possible landmarks for vehicle v
Δ_τ	Time between two predictions
ν_τ^v	Velocity of vehicle v at time τ
ν_τ^v	Differential velocity vehicle v during Δ_τ
$\hat{\nu}_\tau$	Aggregated velocity at time τ
$\hat{\nu}_{\tau-1}$	Aggregated velocity at time $\tau-1$
$\Phi_{\mathfrak{T}_k}$	Centre point of the SDZ of the kth window
$\Phi_{\mathfrak{T}_k,x}$	X coord. of the centre point of the SDZs of kth window
$\Phi_{\mathfrak{T}_k,y}$	Y coord. of the centre point of the SDZs of kth window
$\Phi_{\mathfrak{T}_k,z}$	Z coord. of the centre point of the SDZs of kth window
$\mathfrak{D}(a,b)$	Euclidean distance between point a and point b
$S_{\hat{v}(s,\tau)}$	Aggregated value of sensor s in vehicle v at time τ
$S_{v_u(s,\tau)}$	Sensor s reading reported by user u in vehicle v. time τ
$q_s^{u_v}$	Noise factor of sensor s of user u in vehicle v
$\alpha_s^{u_v}$	Amplification of sensed data s of user u in vehicle v
$h_s^{u_v}$	Fading coeff. of sensor reading s of user u in vehicle v
n_c	Channel noise experienced by the aggregator
$\eta-GPS$	Heterogeneous participant recruitment with GPS
$\eta-nGPS$	Heterogeneous participant recruitment without GPS
$\alpha-GPS$	Active participant recruitment with GPS
$\alpha-nGPS$	Active participant recruitment without GPS

2) Data aggregation: is a fusion center which runs a distributed estimation technique over the acquired data from the sensors. As mentioned above, no adversary behavior is assumed in the data aggregation subsystem but possibly noisy readings and communication channel at each participant that would distort the individual signal transmitted to the aggregator. To this end, the aggregator in this framework simply adopts the estimation technique in [28] with the assumption that transmission of every sensor reading ($S_{v_u(s,\tau)}$) is prone to fading effect of $h_s^{u_v}$, and amplified by $\alpha_s^{u_v}$ whereas the aggregator experiences a channel noise with normalized variance n_c. Given these, the aggregated value of the built-in sensors of U users in vehicle v ($S_{\hat{v}(s,\tau)}$) is formulated as shown in Eq. 1.

$$S_{\hat{v}(s,\tau)} = \sum_{u=1}^{U} (S_{v_u(s,\tau)} + q_s^{u_v}) \cdot \alpha_s^{u_v} \cdot h_s^{u_v} + n_c \tag{1}$$

3) Probabilistic prediction: mainly involves estimating the location of the vehicle within a window of ℓ landmarks, i.e. ***Sensory Detection Zone (SDZ)***. The concept of window of SDZ is illustrated in Fig. 2. Given that the vehicle was predicted to be in a landmark of the timeslot $TS_{k'<k}$ at $t=\tau-\Delta_\tau$, the probability of vehicle v being located in the landmarks of the timeslot T_k at time $t=\tau$ can be calculated as a conditional probability as formulated in as formulated in Eq. 2.

$$\begin{aligned} &Pr(r_\tau^v \in \mathfrak{T}_k^v | r_{\tau-\Delta_t}^v \in \mathfrak{T}_{k'<k}^v) \\ =&\frac{Pr(r_\tau^v \in \mathfrak{T}_k^v \cap r_{\tau-\Delta_t}^v \in \mathfrak{T}_{k'<k}^v)}{Pr(r_{\tau-\Delta_\tau}^v \in \mathfrak{T}_{k'<k}^v)} \end{aligned} \tag{2}$$

Due to the discrete nature of these two occurrences, the probability formulation can be transformed to the simpler form in Eq. 3.

$$\begin{aligned} &Pr(r_\tau^v \in \mathfrak{T}_k^v | r_{\tau-\Delta_t}^v \in \mathfrak{T}_{k'<k}^v) \\ =&\frac{Pr(r_\tau \in \mathfrak{T}_k) Pr(r_{\tau-\Delta_\tau} \in \mathfrak{T}_{k'<k})}{Pr(r_{\tau-\Delta_\tau} \in \mathfrak{T}_{k'<k})} \end{aligned} \tag{3}$$

The probability of a vehicle being in a SDZ of the window $\mathfrak{T}_k$ can be calculated as the ratio of the total distance that can be covered from the centre point of the SDZs of the previous window to the centre point of the SDZs of the kth window ($\mathfrak{D}(\Phi_{\mathfrak{T}_{k-1}}, \Phi_{\mathfrak{T}_k})$) as formulated in Eq. 4. Here it is worth noting that the differential aggregated value of the velocity of the vehicle during the interval of Δ_τ is used to estimate the distance that can be covered. The centre point of window $\mathfrak{T}_k$ is denoted by a vector of xyz coordinates, and and is calculated by obtaining the mean of the x, y, and z coordinates of each SDZ as shown in Eq. 5:

$$\begin{aligned} &Pr(r_\tau \in \mathfrak{T}_k) \\ =&\frac{\sqrt{\mathfrak{D}(\Phi_{\mathfrak{T}_{k-1}}, \Phi_{\mathfrak{T}_k})^2 - \left(\frac{\hat{\nu}_\tau - \hat{\nu}_{\tau-1}}{\Delta_\tau}\right)^2}}{\mathfrak{D}(\Phi_{\mathfrak{T}_{k-1}}, \Phi_{\mathfrak{T}_k})} \end{aligned} \tag{4}$$

$$\begin{aligned} &\Phi_{\mathfrak{T}_k} = [\Phi_{\mathfrak{T}_k,x},\ \Phi_{k,y},\ \Phi_{\mathfrak{T}_k,z}] \\ =&\frac{1}{\ell} \cdot \left[\sum_{i=1}^{\ell} z_{i_x}, \sum_{i=1}^{\ell} z_{i_y},\ \sum_{i=1}^{\ell} z_{i_z} | z_i \in \mathfrak{T}_k \right] \end{aligned} \tag{5}$$

4) Analytics: takes the output of probabilistic estimation of potential landmarks (i.e. SDZs) that the vehicle is expected to be at, and runs (and keeps continuously training) a machine learning-based classifier to estimate the location among the ℓ landmarks.

As mentioned earlier, localizing public transportation vehicles by GPS can be costly in terms of energy, as well as GPS coverage issues. If GPS signals from multiple mobile devices are used to predict the current location on earth, the energy consumption on the participating devices becomes a critical factor as most of those devices run on batteries. Thus, we propose utilizing built-in sensors in smart mobile devices to acquire data, and integrate the sensed data with the proposed framework in Fig. 1 that consists of the four components explained above. The output of the analytics module of the proposed framework is expected to provide an estimated region (i.e. approximate location) of a public transportation vehicle. To this end, while sensor data acquisition application collects data from passengers inside of a public bus, probabilistic prediction module provides an input to the analytics module with a window of predicted stationary regions throughout the static route, which accelerates the classification procedure in the analytics module. The analytics module runs a machine learning algorithm.

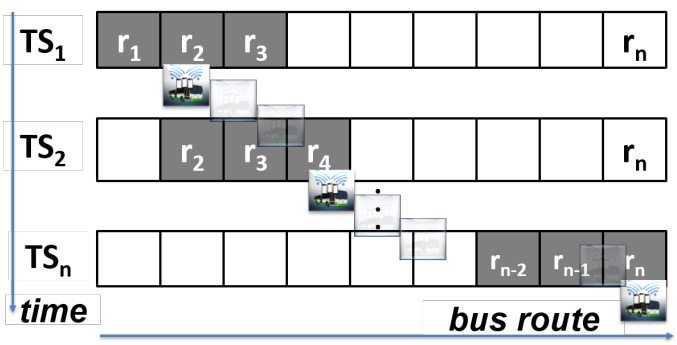

Fig. 2: The concept of sliding window of stationary regions to accelerate the classification process in the Analytics module. It is not possible to estimate the exact location of a vehicle without GPS; however given the probability of a vehicle's presence in a window of ℓ stationary regions (i.e. in this example, $\ell = 3$), probabilistic deduction can be made for the next possible ℓ stationary regions that the vehicle can be present. In the figure, at $\tau = 1$, the vehicle is predicted to be in a stationary region denoted by $\mathfrak{T}_1 = \{r_1, r_2, r_3\}$. If the probability of the vehicle's presence in this window is one, the vehicle is expected to be in $\mathfrak{T}_2 = \{r_2, r_3, r_4\}$ at $\tau = 2$.

Analytics and probabilistic prediction modules run simultaneously. After the initial simulation day, as the analytics module classifies/trains aggregated sensor data, probabilistic prediction module starts making predictions on already classified sensor data with the next simulation day's sensor data.

To clarify the scope of this paper,it is worth mentioning that in this paper, we use Support Vector Machines (SVM) for the purpose of analyzing and making predictions on the acquired data; however feasibility analysis of various machine learning algorithms on the proposed framework is included in the ongoing extensions of this work. Besides, here, sensors that are used to collect data in order to replace GPS during data acquisition process are limited to accelerometer and microphone, whereas various combination of different sensors is also being investigated by our ongoing research agenda.

IV. Performance Evaluation

A. Data acquisition and simulation settings

To evaluate our proposed framework, we have collected real data by using Google's Science Journal application [29] for five consecutive days on a static route under a linear velocity varying between 30-40 kmph towards 15 landmarks (i.e. stops) that are located at least 500 meters apart (available at [30]). The built-in sensors used in data collection are mainly accelerometer (movement along x, y, z axes in terms of meters) and the microphone which senses the sound pressure level in dB. Probability distribution of the collected data has been fed into the CrowdSenSim [31] simulator to generate sensed data by each type of sensor throughout the virtual simulation duration.

TABLE II: Simulation settings

Parameter	Value
Number of vehicles	50
Number of participants	$\{5, 10, 15, 20, 25, 30\}$
Stational region window size (ℓ)	$\{3, 4, 5\}$
Static route length	$\{15, 20, 30\}$
Sensing noise ($q_s^{u_v}$)	$\{1.0, 1.5, 2.0, 2.5, 3.0\}$
Channel noise variance at the aggregator(n_c)	1.0
Average velocity [kmph]	$\{30, 40\}$
Average stop time [s]	$[30, 60]$
Minimum inter-station distance	$0.5km$
Simulation Duration	5 days
Propagation environment	Rayleigh fading
Noise model in sensor readings	AGWN
Classifier	SVM

Simulation settings are summarized in Table II. We have simulated 50 public buses that move over 17 different lines (possibly with overlaps) each of which consists of 15 stops. While any simulation environment can be used to simulate this scenario, we have used Crowdsensim as it already offers data acquisition via mobile crowdsensing.

To simulate bus stops, we have randomly selected 15 random points from the street map in the simulator, which are in form of (latitude, longitude, altitude). Those points are picked according to following criteria: Each consecutive points are ensured to be at least 500 meters apart from each other. There should not be any colliding points. Once the stops (i.e. GPS coordinates of the landmarks) have been determined, they are sorted. As the public transportation vehicles move along static routes, simulating the mobility model is straightforward. Initially each bus is placed on the initial point of its corresponding line. Bus moves forward to its next stop with a constant randomly chosen velocity. After arriving the next stop, the bus waits between 30 seconds to 60 seconds.

This waiting time is also determined randomly on every time a vehicle arrives at a stop. When a bus reaches the final destination, it returns back to its initial position.

In the simulation of the data acquisition module, each passenger's sensor reading is subject to Additive Gaussian White Noise (AGWN). The analytics module uses Support Vector Machine (SVM) with polynomial Kernel type and a Kernel degree of 15 under the termination criteria of either 100 iterations or < 0.01 difference between two consecutive results. In the simulation results, every point represents the average of ten runs with a confidence level of 95%.

B. Numerical Results

We have analyzed the performance of simulation by categorizing our experiments into two sub-experiments in terms of accuracy and energy consumption. Accuracy is calculated as the ratio of the correct predictions to total predictions throughout the total simulation duration (T). A correct prediction denotes the case where the actual GPS coordinates of a public vehicle is within the estimated stational region (r'^{v}_{τ}) as formulated in Eq. 6. We investigate the impact of the following parameters on the accuracy: Window size that is the number of possible landmarks that a vehicle can be present (ℓ), route length, and the noise factor at the built-in sensor reading. All tests have been conducted under varying number of passengers that participate in the crowd-sensing campaign throughout the route.

$$\text{Accuracy} = \frac{\sum_{\tau}^{T} (r^{v}_{\tau} \in TS_{\tau} \wedge r'^{v}_{\tau} = r^{v}_{\tau})}{T} \tag{6}$$

1) GPS-less localization accuracy: Figure 3 illustrates the scenario where route length is set to 15. As seen in the figure, the higher the number of participants, the less the accuracy experienced. This is due to the added noise to the sensor readings as the number of participants (i.e. sensors) increases. On the other hand in each iteration, the larger the window size, the higher the GPS-less localization accuracy. The reason of this phenomenon is that higher number of clusters increases the probability of accurate estimation of the location within the stational regions.

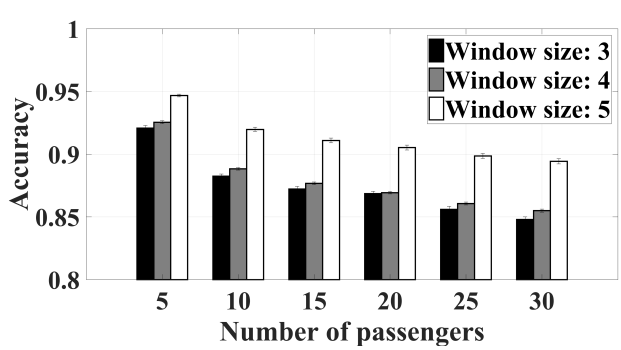

Fig. 3: The effect of window size change on accuracy

As setting $\ell = 5$ results in the best accuracy in the previous figure, we set the window size ℓ to 5 in the second experiment, and vary route length.

As it is illustrated on Fig. 4, longer routes improve the accuracy of GPS-less localization. The best results can be achieved when the number of participants is limited to five and the route length is set to 30 stops. As the window size is large, the noise impacting sensor readings can be low when the number of participants is small, and when the route is long. As the route gets longer, and once the prediction of the next window can be done accurately, classification problem can be done more effectively.

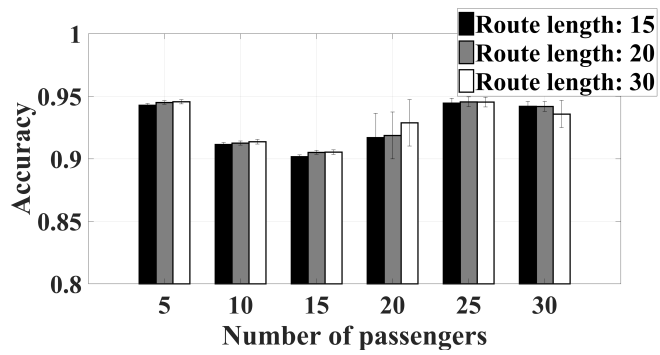

Fig. 4: The effect of route length change on accuracy

To see the effect of AWGN on the accuracy, the third experiment is conducted, the results of which are presented in Fig. 5 where n_i is the variance factor of Gaussian noise. In our simulations, it is the Gaussian distribution of standard deviation of 100 subsequent sensed values of sensor type i. In order to conduct this experiment, standard deviation of 100 subsequent sensors is multiplied with a varying coefficient, n_i. Initial value of n_i is to 1. In every test case, it is increased by 0.5 until 3.0. Here, route length and window size have been set to 5 and 15, respectively as that combination resulted in the best accuracy in the previous results.

As seen in Fig. 5, increasing n_i has a slight effect on the accuracy in most of the test cases except the iterations with 15 and 25 passengers. Nevertheless, the best accuracy has been recorded with 20 passengers.

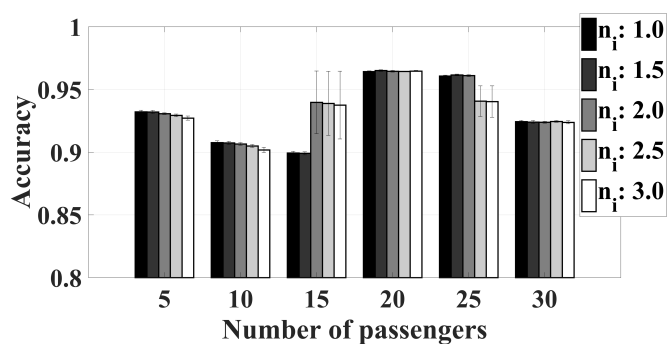

Fig. 5: The effect of Gaussian noise change on accuracy

Based on these results, the following specifications can be made for GPS-less localization in smart public transportation. The overall best accuracy value of 97% can be reached with

the following parameters set: ℓ = 5, route length = 15, number of participating passengers = 20, and n_i = 1.0-3.0. Conversely, the worst overall accuracy has been recorded as 86% with the following specifications: ℓ = 3, route length = 15, number of participating passengers = 30, and n_i = 1.0.

2) GPS-less localization energy consumption: We have compared energy consumption of GPS with the energy consumption of our proposed GPS-less localization framework. To this end, we categorize the passengers/participants into two categories, namely *active* and *inactive*. Activity denotes a coefficient regarding interaction frequency of a user with their mobile device. Energy consumption of an active user (E) in our simulation is the total energy used by previously determined social media applications (E_{apps}) [32] plus the energy consumption of GPS (E_{GPS}) or data acquisition application ($E_{mic} + E_{acc}$) as formulated in Eq. 7.

$$E = \begin{Bmatrix} E_{apps} + E_{mic} + E_{acc} & \eta - nGPS \vee \alpha - nGPS \\ E_{apps} + E_{GPS} & \eta - GPS \vee \alpha - GPS \end{Bmatrix} \quad (7)$$

Under this scenario, passengers are recruited according to following recruitment scenarios:

- $\eta - GPS$
- $\eta - nGPS$
- $\alpha - GPS$
- $\alpha - nGPS$

Heterogeneous participant recruitment acquires data from both active and inactive users whereas active participant recruitment acquires data from only active users. Energy consumption data were obtained from a smart device by applying $\eta - GPS$, $\eta - nGPS$, $\alpha - GPS$, and $\alpha - nGPS$ scenarios. To evaluate energy consumption of a GPS-less participant, Science Journal application kept recoding only accelerometer and microphone data for 30 minutes. On the other hand, GPS was enabled and applications using location services of the smart device kept running in order to obtain energy consumption of a participant using GPS.

Figs. 6.a-b compare the energy consumption of all four recruitment scenarios under best and worst cases in terms of GPS-less localization accuracy. Best case denotes the settings (i.e. route length, number of participants, window length and noise characterization) that results in the best accuracy value whereas the worst case denotes the opposite. As the inactive participant ratio increases in heterogeneous recruitment scenarios, overall energy consumption of mobile device decreases by 38%-46%.

In 6.c-d, multi-dimensional comparison of energy, accuracy and activity is presented under best and worst case settings for localization accuracy of $\eta{-}nGPS$. As seen in both subfigures, increased user activity results in reduced energy consumption due to localization, and better localization accuracy under both settings.

V. Conclusion

Smart transportation and vehicular sensing are crucial components of smart city services. Localization of public transport vehicles by reducing the usage of GPS sensors can be achieved

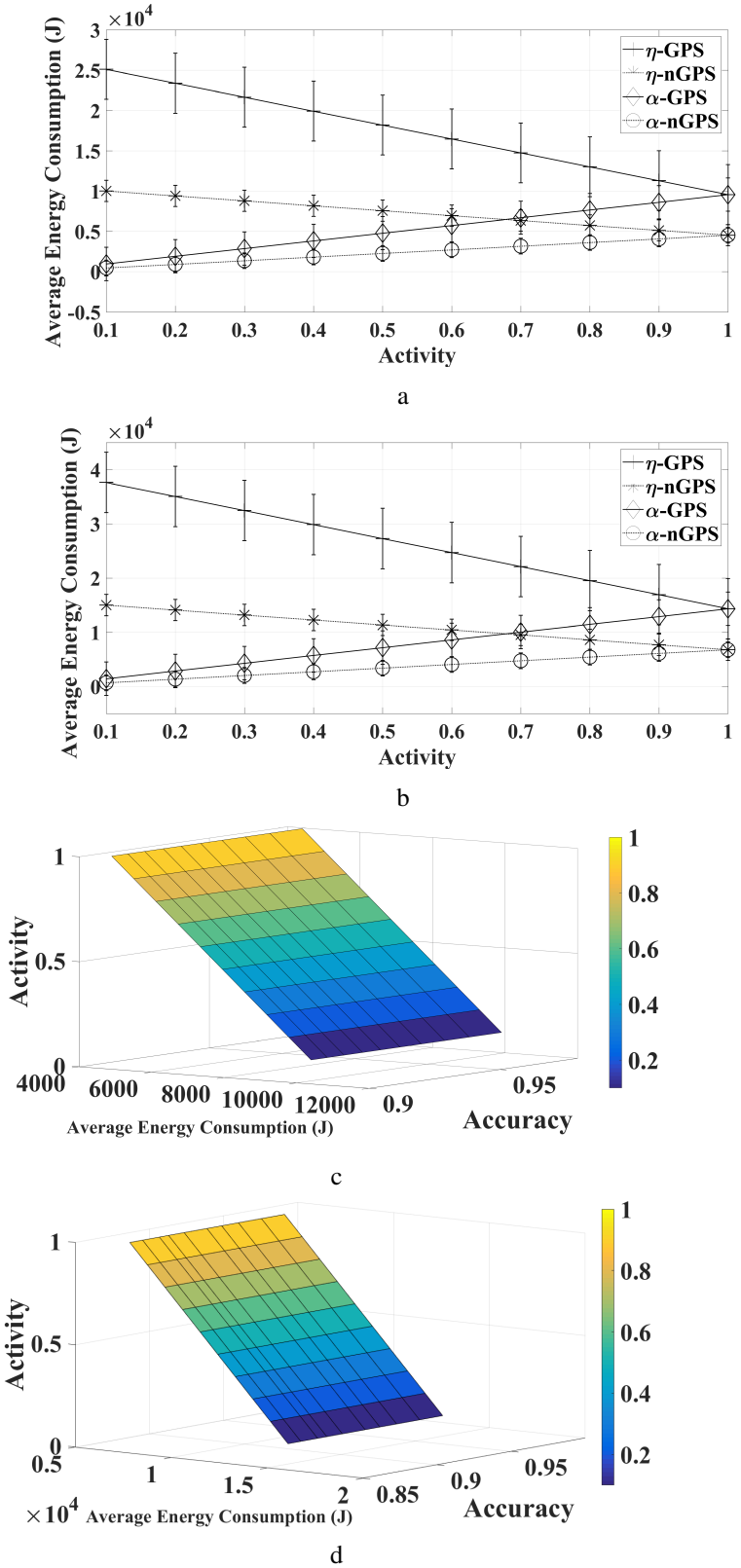

Fig. 6: Energy performance under different scenarios and recruitment schemes (a) Best case energy consumption comparison of $\eta - GPS$, $\eta - nGPS$, $\alpha - GPS$ and $\alpha - nGPS$ (b) Worst case energy consumption comparison of $\eta - GPS$, $\eta{-}nGPS$, $\alpha{-}GPS$ and $\alpha{-}nGPS$ (c) Best case of $\eta{-}nGPS$ energy vs accuracy vs activity (d) Worst case of $\eta - nGPS$ energy vs accuracy vs activity

by covering the location data through other type of sensors that are already available in a non-dedicated manner in a smart infrastructure. Based on this motivation, we have proposed a crowd-sensing-based framework to locate public transportation vehicles in a smart city setting without using GPS. The proposed framework consists of data acquisition, sensory data aggregation, probablistic prediction and analytics modules. Accelerometer and microphones on mobile devices have been used to acquire sensed data and replace GPS. Data collected through a real application have been fed into simulation environment where 50 bus lines have been simulated with each having 15 bus stops. We adopted SVM in the analytics module, and tested the proposed framework by varying the probabilistically derived location region, route length of buses and the characterization of the noise factor introduced from the mobile devices. We have achieved an overall accuracy of above 90%, and 80% in the best and worst cases, respectively. We have also investigated the energy consumption of our best and worst cases performing simulation settings by categorizing participants' active and passive usage profiles. Our results showed that GPS-based localization results in draining up to 46% higher battery power than the overall energy consumption of our proposed GPS-less localization solution.

We are currently testing the proposed framework with different availability of the built-in sensors to investigate the effect of sensor availability on accuracy and energy performance.

Acknowledgment

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under RGPIN/2017-04032 and U.S. National Science Foundation grant CNS-1239423.

References

[1] X. Cheng, C. Chen, W. Zhang, and Y. Yang, "5g-enabled cooperative intelligent vehicular (5genciv) framework: When benz meets marconi," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 53–59, May 2017.

[2] X. Ge, H. Cheng, G. Mao, Y. Yang, and S. Tu, "Vehicular communications for 5g cooperative small-cell networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 7882–7894, Oct 2016.

[3] M. H. Eiza, Q. Ni, and Q. Shi, "Secure and privacy-aware cloud-assisted video reporting service in 5g-enabled vehicular networks," *IEEE Trans. on Vehicular Tech.*, vol. 65, no. 10, pp. 7868–7881, Oct 2016.

[4] J. Harris, M. Beach, A. Nix, and P. Thomas, "The potential of offloading and spectrum sharing for 5g vehicular infotainment," in *2016 IEEE 84th VTC-Fall*, Sept 2016, pp. 1–5.

[5] J. Sahoo, S. Cherkaoui, A. Hafid, and P. K. Sahu, "Dynamic hierarchical aggregation for vehicular sensing," *IEEE Transactions on ITS*, vol. PP, no. 99, pp. 1–18, 2017.

[6] N. A. Ali, M. AbuElkhair, and S. Bouktif, "Utilizing VIN for improved vehicular sensing," in *IEEE WCNC*, April 2016, pp. 1–6.

[7] A. Boukerche, H. A. Oliveira, E. F. Nakamura, and A. A. Loureiro, "Vehicular ad hoc networks: A new challenge for localization-based systems," *Computer Communications*, vol. 31, no. 12, pp. 2838 – 2849, 2008, mobility Protocols for ITS/VANET.

[8] A. Soua and R. Soua, "Suprano: On the localization of connected vehicles in hostile environments: A semi-supervised manifold learning technique," in *2017 IEEE ISCC*, July 2017, pp. 1067–1072.

[9] H. Qin, Y. Peng, and W. Zhang, "Vehicles on rfid: Error-cognitive vehicle localization in gps-less environments," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[10] J. Liu, B. g. Cai, and J. Wang, "Cooperative localization of connected vehicles: Integrating gnss with dsrc using a robust cubature kalman filter," *IEEE Transactions on ITS*, vol. 18/8, pp. 2111–2125, Aug 2017.

[11] M. Pouryazdan, B. Kantarci, T. Soyata, and H. Song, "Anchor-Assisted and Vote-based Trustworthiness Assurance in Smart City Crowdsensing," *IEEE Access*, vol. 4, pp. 529–541, Mar 2016.

[12] M. Pouryazdan, C. Fiandrino, B. Kantarci, T. Soyata, D. Kliazovich, and P. Bouvry, "Intelligent Gaming for Mobile Crowd-Sensing Participants to Acquire Trustworthy Big Data in the Internet of Things," *IEEE Access*, vol. 5, no. 1, pp. 22 209–22 223, Dec 2017.

[13] C. X. Mavromoustakis, G. Mastorakis, J. M. Batalla, and P. Chatzimisios, "Social-oriented mobile cloud offload processing with delay constraints for efficient energy conservation," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–7.

[14] M. A. Demetriou, "Emulating a mobile spatially distributed sensor by mobile pointwise sensors in state estimation of partial differential equations via spatial interpolation," in *2017 American Control Conference (ACC)*, May 2017, pp. 3243–3248.

[15] J. Boomgaarden, J. Corney, H. Whittaker, G. Dinolt, and J. McEachen, "Challenges in emulating sensor and resource-based state changes for android malware detection," in *9th Intl. ICSPCS*, Dec 2015, pp. 1–10.

[16] A. Boukerche, H. A. B. F. Oliveira, E. F. Nakamura, and A. A. F. Loureiro, "Dv-loc: a scalable localization protocol using voronoi diagrams for wireless sensor networks," *IEEE Wireless Communications*, vol. 16, no. 2, pp. 50–55, April 2009.

[17] H. A. B. F. de Oliveira, A. Boukerche, E. F. Nakamura, and A. A. F. Loureiro, "An efficient directed localization recursion protocol for wireless sensor networks," *IEEE Transactions on Computers*, vol. 58, no. 5, pp. 677–691, May 2009.

[18] H. A. Oliveira, A. Boukerche, E. F. Nakamura, and A. A. Loureiro, "Localization in time and space for wireless sensor networks: An efficient and lightweight algorithm," *Performance Evaluation*, vol. 66, no. 3, pp. 209 – 222, 2009, modeling and Analysis of Wireless Networks: Selected Papers from MSWiM 2007.

[19] Y. Liu, X. Weng, J. Wan, X. Yue, H. Song, and A. V. Vasilakos, "Exploring data validity in transportation systems for smart cities," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 26–33, May 2017.

[20] J. Jung and K. Sohn, "Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data," *IET Intelligent Transport Systems*, vol. 11, no. 6, pp. 334–339, 2017.

[21] P. Ta-Shma, A. Akbar, G. Gerson-Golan, G. Hadash, F. Carrez, and K. Moessner, "An ingestion and analytics architecture for iot applied to smart city use cases," *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–1, 2017.

[22] R. Jia, P. Jiang, L. Liu, L. Cui, and Y. Shi, "Data driven congestion trends prediction of urban transportation," *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–1, 2017.

[23] F. Zhu, Z. Li, S. Chen, and G. Xiong, "Parallel transportation management and control system and its applications in building smart cities," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1576–1585, June 2016.

[24] M. Aloqaily, B. Kantarci, and H. T. Mouftah, "Multiagent/multiobjective interaction game system for service provisioning in vehicular cloud," *IEEE Access*, vol. 4, pp. 3153–3168, 2016.

[25] D. Liao, G. Sun, M. Zhang, V. Chang, and H. Li, "Towards location and trajectory privacy preservation in 5g vehicular social network," in *IEEE Intl. Conf. on Computational Science and Engineering and IEEE Intl. Conf. on Embedded and Ubiquitous Computing*, vol. 2, July 2017, pp. 63–69.

[26] A. Jahangiri and H. A. Rakha, "Applying machine learning techniques to transportation mode recognition using mobile phone sensor data," *IEEE Trans. on ITS*, vol. 16, no. 5, pp. 2406–2417, Oct 2015.

[27] M. Pouryazdan, B. Kantarci, T. Soyata, L. Foschini, and H. Song, "Quantifying User Reputation Scores, Data Trustworthiness, and User Incentives in Mobile Crowd-Sensing," *IEEE Access*, vol. 5, pp. 1382–1397, Jan 2017.

[28] M. K. Banavar, C. Tepedelenlioglu, and A. Spanias, "Estimation over fading channels with limited feedback using distributed sensing," *IEEE Trans. on Signal Processing*, vol. 58, no. 1, pp. 414–425, Jan 2010.

[29] Google, "Science Journal," https://makingscience.withgoogle.com/science-journal.

[30] Nextcon Lab, "Dataset of All Experiments," http://nextconlab.academy/ICC/.

[31] C. Fiandrino, A. Capponi, G. Cacciatore, D. Kliazovich, U. Sorger, P. Bouvry, B. Kantarci, F. Granelli, and S. Giordano, "Crowdsensim: a simulation platform for mobile crowdsensing in realistic urban environments," *IEEE Access*, vol. 5, pp. 3490–3503, 2017.

[32] F. Anjomshoa, M. Aloqaily, B. Kantarci, M. Erol-Kantarci, and S. Schuckers, "Social behaviometrics for personalized devices in the internet of things era," *IEEE Access*, vol. 5, pp. 12 199–12 213, 2017.