

Vehicular dataset for road assessment conditions

Mário Antunes, Diogo Gomes, João Paulo Barraca and Rui L. Aguiar

Instituto de Telecomunicações

Universidade de Aveiro

Aveiro, Portugal

Email: mario.antunes,dgomes,jpbarraca,ruilaa@av.it.pt

Abstract—The Internet of Things (IoT) is a very promising concept that by connecting numerous devices to the internet and extracting large sums of information (BigData) can enable the realisation of various futuristic scenarios. In order to develop and assess future applications and services, it is necessary the availability of datasets that can be used to train, test and cross validate. Project SCoT (Smart Cloud of Things) has developed an M2M platform capable of collecting information from heterogeneous devices and collide that information in a large data repository. During its pilot phase, the project made the assessment of the road conditions in the region of Aveiro, Portugal. In this work we make the dataset used on the previous mentioned pilot publicly available. With this dataset our road assessment algorithm reached 80% accuracy in the task of pothole detection, other scenarios (that take into account vehicular speed, position and acceleration) can also be explored. The dataset was not pre-processed in anyway, the only transformation was made to protect the identity of the volunteers.

Index Terms—IoT, M2M, Machine Learning, Dataset

I. INTRODUCTION

Over the last years the Internet of Things (IoT) [1] has gained significant attention from both industry and academia. IoT has made it possible for everyday devices to acquire and store contextual data, in order to use it at a later stage. This allows devices to share data with one another, and even services on the Internet in order to cooperate and accomplish a given objective. A cornerstone to this connectivity landscape is machine-to-machine (M2M) [2]. M2M generally refers to information and communication technologies able to measure, deliver, digest and react upon information autonomously, i.e. with none or minimal human interaction.

The data generated by these devices are an untapped source of context information. This information can be used to provide added value: improve efficiency, detect abnormal conditions or advertise information. As microcosms of IoT, cities stand to benefit the most from the untapped information shared by all these devices. Smart cities means many things to numerous people. Yet, one thing remains constant: part of being “smart” is utilising information and communications technology and the Internet to address urban challenges. Fusing information from several sensors

makes it possible to predict a driver’s ideal parking spot [3], [4]. Projects such as Pothole Patrol [5] and Nericell [6] use vehicular accelerations to monitor road conditions and detect potholes. TIME (Transport Information Monitoring Environment) project [7] combines data from mobile and fixed sensors in order to evaluate road congestion in real time.

In this paper we present a publicly available dataset collected from the outcomes of project SCoT. The focus of this paper is to make the road condition assessment dataset public available for further research. Although the dataset was primarily used to detect potholes in the road pavement, it can be used in other scenarios that take into account vehicular acceleration, speed and position. The remainder of the paper is organized as follows. In Section II we briefly describe the SCoT platform. The road assessment pilot is described in detail in Section III. All the important details about the dataset are presented in Section IV. Finally, the conclusion remarks are given in Section V.

II. SMART CLOUD OF THINGS (SCoT)

The SCoT platform [8] is an evolution over our previous work, APOLLO [9], aiming at the development of a generic platform for integration of IoT/IoS (Internet of Services) scenarios. Like its predecessor, it covers aspects related to network, device management, services and applications overcoming the shortcomings of the solutions previously identified, and presenting novel data mining concepts. An important aspect is that we considered the entire M2M ecosystem, and its stakeholders.

We assume that an existing Telecommunication provider infrastructure, and their Operation Support Systems (OSS) is present and capable of being fully integrated, through the standardised interfaces. Such integration is desirable as OSS provides many of the desirable functionality and enables full integration of our platform into an existing environment. SCoT aims to allow multiple tenants to deploy their services with agility and reduced time to market, over a wide range of scenarios and using different sensors.

The platform abides to ETSI M2M and can be divided in four major domains: Sensor, Network, Service, and Data (see Figure 1). These domains are closely related with IoT/IoS, enabling the Telecommunication operator to act as the

vital glue holding both concepts together, and presenting an offer with added value to its clients.

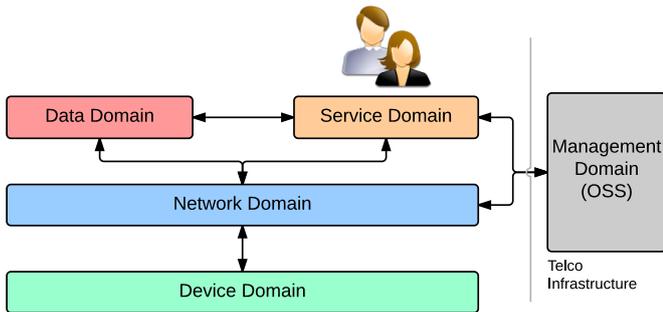


Fig. 1. Architecture of the SCoT platform.

III. ROAD CONDITION ASSESSMENT

We targeted road condition assessment through pothole detection, recurring to crowd-sourcing, massive data collection, using off-the-shelf mobile devices and machine learning techniques. An Android App was created and made available to citizens who would place their monitoring phones in the dashboard of their cars.

Each monitoring phone would monitor the location, speed, and 3 axis acceleration with a frequency of 15Hz. The system assesses the road surface condition of several vehicles (use case similar to [5]). Sensors report information every 5 hours using their 3G connection, or immediately if a Wifi connection was available. Data flows to an intermediate gateway, and then is dispatched to the components at the network layer. Finally, information is stored in several databases for the purpose of benchmarking, analysis and context enrichment.

The documents generated by the vehicles are filtered in order to detect high peaks in the Z (vertical) axis. After we leveraged our cluster based storage for detecting anomalies based on high Z peaks events, and a machine learning approach for determining anomalies based on a reference road segment. The model is depicted in Figure 2, we collected a validation dataset composed of 216 potholes around University of Aveiro and use it to train our model using a genetic algorithm for parameter selection. In our implementation we used a Global Parallel Genetic Algorithm (GPGA), a genetic algorithm that computes the fitness of all the chromosomes in parallel with tournament selection to select the progenitors for the next generation. In Table I we details the main parameters of our implementation.

Our model is based on a set of filters, similar to [5]:

Z-filter: filters out events with z peaks smaller than a given threshold. Main indication of a pothole or bump in the road pavement.

YZ-filter: filters out events with a ration Y/Z peak smaller than a given threshold. Filter out speed bumps that span across the lane.

TABLE I
MAIN PARAMETERS OF THE GENETIC OPTIMIZATION

Parameter	Value
Population Size	250
Number of iterations	1000
Mutation Probability	20%
Tournament Size	3

Speed/Z-filter: filters out events with a ration $Speed/Z$ peak smaller than a given threshold. Filter out accelerations provoke by higher speeds.

After the filtering process, the pothole candidates are grouped together with a fast greedy clustering algorithm. Instead of relying on a conventional clustering method, we devised a fast greedy algorithm, for two main reasons. First, processing speed was an important requirement of the described platform. Second, it is difficult to estimate the correct number of clusters, instead we known the average precision of the GPS. Taking this value into account we can control the maximum radius of each clusters. The final result may be a local optimum, however it is more than enough to implement a voting system. The algorithm uses KD-Trees to find the closest pair of points and merges them together while the cluster maintains an inter-cluster distance smaller than a certain threshold (in this case was the average precision of the GPS). Algorithm 1 describes in detail the inner-workings of the method. A KD-Tree was used to speed up the process of finding the closest pairs. A naive approach uses $\mathcal{O}(n^2)$ operations to computed the distances and $\mathcal{O}(n^2)$ to filter out the closest pairs, witch ends up as $\mathcal{O}(n^2)$. On the other hand, a KD-Tree uses $\mathcal{O}(n \log n)$ operations to build a tree, and another $\mathcal{O}(n \log n)$ to find the closest pairs of all the points. Ending up as with a complexity of $\mathcal{O}(n \log n)$, which is significantly better then $\mathcal{O}(n^2)$.

As a result we obtained 82% in determining potholes, under realistic conditions. We had no control over the vehicle, driving style, vehicle condition, or cell phone location. We processed tens of million reports per month, which enabled us to build a detailed map covering the entire Aveiro municipal region, and even part of the centre region of Portugal.

IV. DATASET

In this section we describe in greater detail the dataset that was gathered during the project. Although the dataset was primarily used for pothole detection (achieving a high accuracy in that task) it can be used for other tasks/scenarios related with vehicular position, acceleration and speed. In Figure 3 we depict a graphical representation of the dataset.

Algorithm 1 Fast Greedy Clustering Algorithm

```
1: function CLUSTERING(points)
2:   kdTree  $\leftarrow$  KDTree.init(points, radius)
3:   pairs  $\leftarrow$  {}
4:   for each p  $\in$  points do
5:     neighbours  $\leftarrow$  kdTree.near(p, radius)
6:     pairs  $\leftarrow$  pairs + neighbours
7:   end for
8:   clusters  $\leftarrow$  {}
9:   for each p  $\in$  pairs do
10:    c  $\leftarrow$  findClosestCluster(clusters, p)
11:    if maxRadius(c + p) < radius then
12:      c  $\leftarrow$  c + p
13:    end if
14:  end for
15:  return clusters
16: end function
```

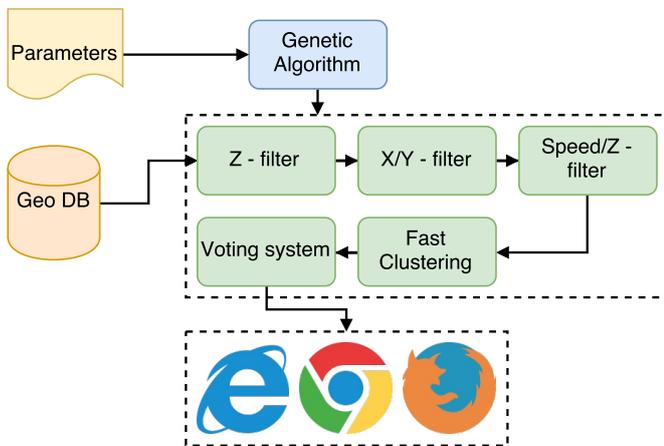


Fig. 2. Block model of the road assessment model and learning scheme.

The dataset is public available¹. It is stored in JSON format, divided into several files (with two different schemas): i) dataset[**date**].json contains a portion of the dataset collected by volunteers for that respective **date**, and ii) results.json contains the pothole validation set collected manually.

In JSON 1 we describe the schema of the dataset[**date**].json files. Each monitoring phone captured location, speed, and 3 axis acceleration with a frequency of 15Hz. For each GPS position exists a list of acceleration values. The list does not have a fixed size, it depends on the phone hardware and software. We also gathered the timestamp of each acceleration measure, this improves the computation of first and second derivatives. Ideally the Z axis corresponds to up and down; Y axis corresponds to left and right and finally the X axis corresponds to front and back. However, this ultimately depends on

¹<https://atnog.av.it.pt/mantunes/road>

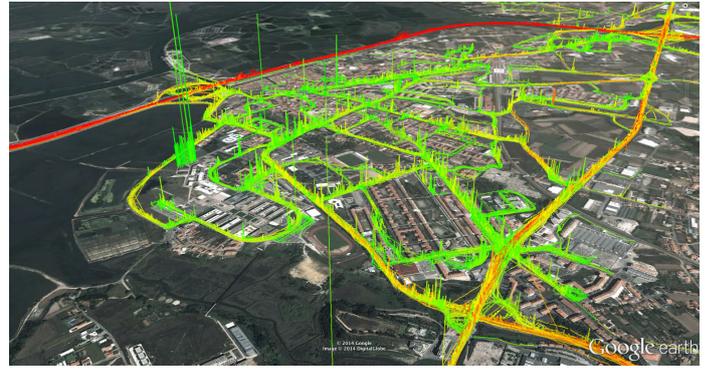


Fig. 3. Graphical representation of the dataset. The color represents velocity (green to slower speeds and red to higher speeds). The height of the line represents the acceleration on the Z axis.

the volunteer position of the phone. Even when the phone was positioned correctly we have to consider the dashboard inclination, and movement during the trip. The dataset is not pre-processed in any way. We only applied a single transformation to remove certain initial parts of the trip and randomised the id to protect the volunteers' identity. A tool used to anonymise the dataset is publicly available², it detects individual trips, generates a random identifier and removes the initial and final part (3 minutes) of the trip to minimize the possibility of fingerprinting. Another important parameter in the dataset is the version field, the android application used to capture the dataset was improved during the project execution, with various improvements and regressions along its way, the version field captures those variances in the captured data. Still, we cannot guarantee which version was used by the volunteers, but it is possible to filter the dataset for a specific version. The speed field contains the speed estimated by the GPS movement, may not correspond to the real speed of the vehicle.

JSON 1. Schema for dataset.json

```
{
  "dataset": [{
    "id": // unique id for each trip
    "version": // version of the app
    "latitude": // latitude of the event
    "longitude": // longitude of the event
    "speed": // speed of the vehicle
    "accelerometer": {
      "x": [] // acceleration on the x axis
      "y": [] // acceleration on the y axis
      "z": [] // acceleration on the z axis
    }
    "timestamp": []
    // timestamps of the accelerations
  }
  ...
}
```

²<https://github.com/ATNoG/road-dataset-anonymization>

]]

In JSON 2 we describe the schema of the results.json files. This file contains the location of several potholes around the University of Aveiro. The potholes were classified into 7 classes:

Pothole small: a single pothole smaller than 20 cm in diameter

Pothole large: a single pothole larger than 20 cm in diameter

Pothole cluster: multiple potholes clustered together

Drain pit: drain pit at the side of the road

Gap: a large gap that span at least a lane

Bump: bump on the road pavement

Speed bump: bump on the road pavement, with the objective of reducing speed

We also included two geographical points to define the minimum-bounding box. Anyone that intends to use the results.json file can filter out events that occurred outside the bounding box. It is also important to notice that these datasets have a validity period. The road pavement was repaired and damaged, after the data acquisition, as a result this file contains a validity time period.

JSON 2. Schema for results.json

```
{
  "date": {
    "begin":
    "end" :
  },
  // Period in witch the dataset is valid
  "minimun bounding box" : [{
    "latitude":,
    "longititude:"
  }...],
  // Geographical positions that define
  // the bounding box of the results
  "results": [{
    "event":
    // Type of pothole: pothole_small;
    //pothole_large; pothole_cluster;
    //drain_pit; gap; bump; speed_bump
    "latitude": // latitude of the event
    "longititude": // longititude of the event
  }...]}]
```

V. CONCLUSIONS

In this paper we publicly share the dataset captured during projecto SCoT and used to validate its data processing capabilities through pothole detection This dataset was used to train a road assessment algorithm with high accuracy, but can be used for any other scenario involving vehicular acceleration, speed and position. In this paper we make public available (under Open Source Licenses)the road assessment dataset for further research in the hope that other researchers might mind this dataset useful in new studies.

ACKNOWLEDGEMENT

This work is supported by the European Structural Investment Funds (ESIF), through the Regional Operational Programme of Centre (CENTRO 2020) [Project Nr. CENTRO-01-0246-FEDER-000008].

REFERENCES

- [1] F. Wortmann, K. Flüchter *et al.*, "Internet of things," *Business & Information Systems Engineering*, vol. 57, no. 3, pp. 221–224, 2015.
- [2] K.-C. Chen and S.-Y. Lien, "Machine-to-machine communications: Technologies and challenges," *Ad Hoc Networks*, vol. 18, pp. 3–23, 2014.
- [3] T. Rajabioun, B. Foster, and P. Ioannou, "Intelligent parking assist," in *Control Automation (MED), 2013 21st Mediterranean Conference on*, June 2013, pp. 1156–1161.
- [4] J. K. Suhr and H. G. Jung, "Sensor fusion-based vacant parking slot detection and tracking," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 15, no. 1, pp. 21–36, February 2014.
- [5] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, "The pothole patrol: Using a mobile sensor network for road surface monitoring," in *Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '08. New York, NY, USA: ACM, 2008, pp. 29–39. [Online]. Available: <http://doi.acm.org/10.1145/1378600.1378605>
- [6] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones," in *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '08. New York, NY, USA: ACM, 2008, pp. 323–336. [Online]. Available: <http://doi.acm.org/10.1145/1460412.1460444>
- [7] J. Bacon, A. Bejan, A. Beresford, D. Evans, R. Gibbens, and K. Moody, "Using real-time road traffic data to evaluate congestion," in *Dependable and Historic Computing*, ser. Lecture Notes in Computer Science, C. Jones and J. Lloyd, Eds. Springer Berlin Heidelberg, 2011, vol. 6875, pp. 93–117.
- [8] M. Antunes, J. P. Barraca, D. Gomes, P. Oliveira, and R. L. Aguiar, "Smart cloud of things: an evolved iot platform for telco providers," *Journal of Ambient Wireless Communications and Smart Environments (AMBIENTCOM)*, vol. 1, no. 1, pp. 1–24, 2015.
- [9] —, *Unified Platform for M2M Telco Providers*. Cham: Springer International Publishing, 2014, pp. 436–443. [Online]. Available: https://doi.org/10.1007/978-3-319-13102-3_71