Visualization System for Human Mobility Analysis

Adriana Drăghici, Taygun Agiali, Cristian Chilipirea
Computer Science and Engineering Department
University Politehnica of Bucharest, Romania
{adriana.draghici, cristian.chilipirea}@cs.pub.ro, taygun.agiali@cti.pub.ro,

***Abstract* — City-scale events, such as festivals, sport events, concerts gather thousands of people in different areas of a city. By installing large area Wi-Fi routers which record the presence of people using devices such as smartphones and tables, the density and the movements of the crowds can be estimated and classified. Analysis of crowd data collected during such events leads to valuable insights and feedback both for the parties involved and for city planners and advertisers. In this paper we present a analysis and visualization system for improved understanding of density and movement information. We discuss real datasets constraints, data clean-up techniques and methods for estimating and classifying the crowd movements. These methods are implemented inside a framework that allows easy integration and analysis of mobility datasets. As proof of concept we present the analysis of a real dataset collected during a city-scale experiment in Arnhem, the Netherlands.**

***Keywords* — pedestrian mobility; crowd density estimation; Wi-Fi hotspots;**

I. Introduction

The overgrowth of city populations [2] places a big stress on their infrastructure and their capacity of handling high density crowds of pedestrians. In smaller cities this kind of stress can be induced by tourism and city-scale events that bring a large number of visitors in a short interval in a limited space environment. In Europe there are many festivals organized in smaller cities that attract a number of participants larger than the population.

The evolution of hardware and processing power and the fact that portable devices such as smartphones and tablets have become ubiquitous [4], leads to the implementation of various pedestrian and traffic data collection systems. Such systems provide large sets of data concerning the flow and density of people throughout the city or in closed environments (parks, stadiums, train stations etc) by collecting the packets sent regularly by the Wi-Fi modules embedded in their devices. The characteristics and accuracy of the dataset depends in a great amount on the method used for collection and also on the type of environment.

We propose an analysis system for pedestrian mobility data that interested parties (such as event organizers and local authorities) may use to better grasp the city's pulse. The input datasets are collected using Wi-Fi hotspots that record the presence of adjacent wireless devices. Related projects (see section II) deal mostly with Bluetooth or GPS collected data, most of them in closed open air environments, not city scale, and their analysis is particular to those sets of values. Our study focuses on outdoor city-wide environments, with emphasis on visualization methods, and is extensible, i.e. not customized for a certain experiment.

Our contributions include proposals of scenarios and techniques for movement and density analysis of human mobility throughout a city. We show the corresponding graphical representations that best express the results. We validate our system on a real dataset collected in Arnhem, The Netherlands during a three weeks period that also included recordings for a weekend festival.

The paper is structured as follows. First, we outline the recent research on the topic of crowd monitoring. In Section 0 we describe our design choices and the characteristics of our system while in Section **Eroare! Fără sursă de referință.** we present our analysis approach and its results and visualizations for the Arnhem dataset. Section V offers conclusions and possible directions of future work.

II. RelAted Work

The research of crowd dynamics is an evolving topic and even though many studies focused at first on identifying or even preventing critical situations [8] [9] we now see a tendency to also consider the quality of experience during large-scale events.

The rise of the mobile devices industry led to powerful smartphones equipped with a multitude of sensors for localization and motion and also support for communication protocols such as Bluetooth, NFC or Wi-Fi Direct. The common practice is to obtain the device location via a custom application, but Nishimura et al [11] offered a more interesting approach that uses the device accelerometer and audio recordings to classify the congestion of public spaces.

Bluetooth is a popular technology for crowd sensing, but has certain limitations when applied in large open environments. Moreover studies have shown that if we want to detect devices without requiring participants to use special tags or apps, we won't be successful. Schauer et al [12] conducted an experiment in a crowded airport by scanning for Bluetooth and Wi-Fi Devices, and have shown that Wi-Fi offers a good approximation of the actual densities and flow, while Bluetooth was much less accurate.

Blanke et al. [6], Wirz et al. [14] and Stopczynski et al. [13] analyzed crowd dynamics during large gatherings using participatory applications. In such contexts, the biggest problem is how to attract a larger number of users, and only Blanke et al. addressed this situation by providing an official event app with

978-1-4673-8180-2/15/$31.00 ©2015 IEEE

social media incentives incorporated in it. As for the analysis techniques, Wirz et al. employ heat maps and calculate the crowd velocity, Stopczynski et al. are interested in the spatial coverage during a festival and Blanke et al. make a more thorough analysis by visualizing not only crowd density but also crowd flow. Even though they worked with finer grained data than we expect from Wi-Fi hotpots based collection systems, their methods can be applied to such datasets. What we do different is shift the focus from the collection process to the analysis process; we incorporate such techniques in a framework and permit them to be applied on various hotspot based experiments.

III. System Design

As discussed in chapter II, several technologies can be employed for crowd sensing, and in the past two years the tendency is to monitor large open environments using GPS traces, Bluetooth or Wi-Fi hotspots detections. Our work focuses on the latter and provides an analysis and visualization system customized for mobility data collected from such areas. To our knowledge, at the time we started its implementation it was the first one dedicated to this kind of analysis that was not experiment dependent. We designed a software framework that allows us to incorporate analysis methods and apply them on any given mobility dataset that respects a set of characteristics (provides addresses, timestamps, collection points (hotspots)). We integrated methods and heuristics customized for mobility data; it has three main layers, one for cleaning and filtering, one for analysis and one for visualization, as presented in figure 1. We incorporated features such as *experiment-independence, extensibility and mitigation of data errors*.

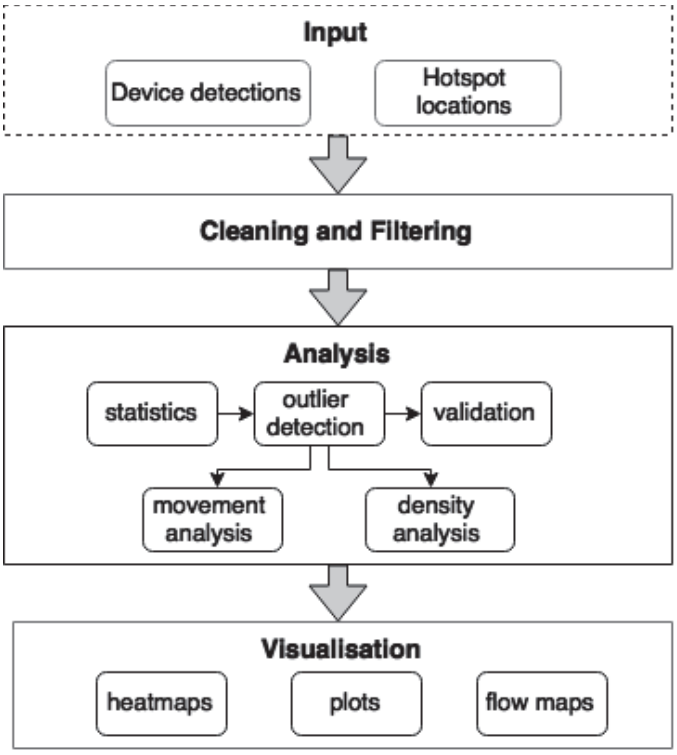

Fig. 1. System layers, we import data from various formats, store it and perform basic cleanup in the first layer, we have several components in the analysis layer, some depending on the output of others and we provide several visualization methods in the last layer

The infrastructure that provides data consists of a set of Wi-Fi routers (hotspots) placed in well-studied positions throughout a large area of a city. These hotspots detect Wi-Fi devices, obtaining their MAC address. In order to preserve privacy, they store and transmit to a server the hash of the address. The server receives these detections and stores them. The infrastructures we have received datasets from, used relational databases and provided us with SQL dumps and comma-separated value files. Along with the hash, these systems record timestamps and hotspot related data. The infrastructure for the data discussed in chapter **Eroare! Fără sursă de referință.** did not provide signal strength data (RSSI), but our system supports it.

Dataset Characteristics

The analysis framework uses a simple data format which is composed of the information regarding the routers and the records of device detections. Concerning the actual detections a minimal set of attributes should be recorded: the ID of the hotspot that collected the detection, an identifier for the device, in the literature we found this is common to be the MAC address of the device, and the timestamp of the detection. To increase the spatial granularity of the detections and to allow for a better localization, the monitoring system may also record the maximum and median signal strength of each detection. The first dataset we analyzed did not provide these signal related values; nonetheless, we included support for them in our system, since other datasets might provide them.

Cleaning and Filtering Layer

When working with real large datasets one must also consider the quality of entries. The cleaning layer represents the first stage towards the extraction of useful information from gathered data. This layer is intended to detect, correct or remove corrupt and inaccurate entries from the database, while keeping the data consistent with other similar data sets. Erroneous entries could also point out to collecting system design issues and software implementations bugs. The correction of these errors improves the methodology used in collecting device detections. In order to assess the sanity of the dataset, each attribute of a detection record is type checked to avoid entries with missing or misleading information: the range of timestamps, the validity of identifiers etc. Entries that do not satisfy these constraints are removed from our dataset. In the implementation, this layer also offers a component for importing datasets in various formats, storing them into a database and another component that mediates access to data.

The outliers detection and removal and the identification of the Ping Pong effect as part of this layer; since it required more in depth analysis than a simple verification of records, we implemented it in the analysis layer.

Analysis Layer

The task of the Analysis Layer is to perform analysis on cleaned data, out of which crowd characteristics such as density and movement could be deducted. Most of the output of this layer is input to the components of the visualization layer.

The theoretical domain for this kind of analysis is referred to as Episodic Movement Data Analysis and thoroughly described by Thomas Liebig [10]. Applying this theory, the

objects of our analysis are described as sets of records with the following attributes: object identifier o_k (in our case the device id), spatial position p_i (hotspot location), time t (detection timestamp) and possibly other attributes (e.g. signal strength). On the given dataset we use these attributes to formally define and identify *visits* and *moves*. These are the basic elements for our analysis and using them we look compute *time series associated with places, time series associated with links between hotspots* and *spatial flow* for individual devices and for several. We use these for capturing the crowd density and the crowd movements throughout the city.

Detection and removal of *outliers* in a dataset improve the accuracy of the proposed analysis. Depending on the size of the dataset and the accuracy of the collecting methodology, the extraction of outliers would decrease the quantity of the dataset, to the point that is not sufficient for further analysis. Using statistics about the detections we can detect when hotspots did not function as intended and also identify and exclude devices that have too many detections (e.g. one device has 40% of total number of detections). The resulting dataset is used as input for the rest of our analysis.

Because we work with data collected from Wi-Fi hotspots and not with app-collected GPS locations like other existing systems, we can encounter the *Ping Pong effect*, i.e. the detection of one device by multiple hotspots. The solution we implemented in our system is to create a virtual hotspot in the intersection of the overlapping areas by removing the existing detections for those hotspots and creating detections for the newly formed hotspot.

We included a module for *validation* that can be extended with various heuristics. Since we usually have no other recordings regarding pedestrians movement throughout the city we decided to look for *day-night patterns* and to compare *weekend densities*. Researchers can combine the output of this module with domain knowledge to make some assumptions about the patterns we observe, such as events that may take place in some weekends or evenings and characteristics of the areas (train stations, commercial venues, offices etc.).

The modules for crowd density and crowd movements provide the input for the visualization layer.

Visualization Layer

The purpose of the Visualization Layer is to represent empirical findings as graphs, heat maps, charts and tables. This layer is vital to the overall analysis because it allows both researchers and domain experts to get a better grasp of the data. This layer exposes its characteristics and allows comparisons between different space and time locations. With information presented this way, the user understands the causality of some cycled actions, which conclude in patterns that can predict further states of the dataset. In order to visualize empirical findings in our data set, this layer uses line and scatter *plots, dynamic heatmaps, flowmaps*, bar, pie and bubble *charts*. The Visualization Layer works in correlation with the Cleaning and Analysis Layer and is accessible using the system's web interface or its API. Regarding the visualizations performed on cleaned data, the charts ensure their validity by presenting similar trends between different datasets (comparing the density values for different week-ends or different days) and also validates known patterns, such as the citizens' diurnal and nocturnal activity.

IV. Analysis of a Real World Trace

We began our analysis and system design by determining what are we really looking for in a crowd dataset, and especially in one covering a large area of a city. Our system has to capture crowd dynamics in a way that benefits interested parties such as event organizers, advertisers or city officials. In this context we emphasize the role of visualization and we looked for the best ways to present and model the quantities and movements of pedestrians.

One of the design principles of our system is its experiment independence, meaning that we can use input datasets from different hotspot-based sensing infrastructures and use its interface to obtain various statistics and visualizations. Depending on the experiment, one can even extend the existing API with customized methods or heuristics.

Our system is designed to work with datasets ranging from a few days to a few weeks or even months, databases that may have different analysis requirements. To implement such a design we took into consideration several possible *scenarios* for which our framework offers visualizations: *Festival, Popular Areas, Location Habits*. In the case of a *festival* with several attractions inside the city or nearby (e.g. camping), the event organizers are interested in the popularity of its attractions, but also in the usage of public facilities or other city attractions and locations such as restaurants and train stations. The analysis of such data should focus on determining the quality of experience during the event. City-scale festivals usually last from a few hours to three or four days, so we do not look in this data for daily mobility patterns but focus on assessing the densities around the attractions and the flow between them. We are also interested in the duration of visits (the time in which a device is detected by a hotspot) and how much each route between the attractions is taken. For the *Popular Areas scenario* we need to capture people's tendency to visit certain locations, and as future work we can combine the hotpots detection with social media data such as Fourquare traces [1]. In this scenario we focus less on moves but on visits and look for density patterns throughout several days. In the *Location Habits scenario* we include in our model the nature of that location (office, shop, station etc) and look for the tendencies of individuals to stay or to move between them. Here the focus shifts from area density to the movement analysis.

To support these scenarios, our Analysis Layer can output data regarding:

- walking speeds and transportation classification (on foot, on bicycle)
- moves between hotspots, individual devices routes, link densities
- visit counts and visits duration
- detection counts

A. Arnhem Dataset Evaluation

Data used to evaluate our methodology was collected using Wi-Fi hotspots placed in Arnhem, The Netherlands, a city with a population of 152,850, between 27 September and 18 October 2014. This infrastructure was deployed as part of the project EWiDS (Extreme Wireless Distributed Systems) [3] and is detailed in [7]. Sixteen Wi-Fi hotspots were placed such as to cover the city center area, as depicted in Figure 2. The privacy of detected devices was ensured by using a hashing function to

encode the device MAC address, installed in every software of Wi-Fi hotspots. We have split the recordings into two datasets, one corresponding to an art festival talking place during the 27th and 28th September and the other for the days that followed it.

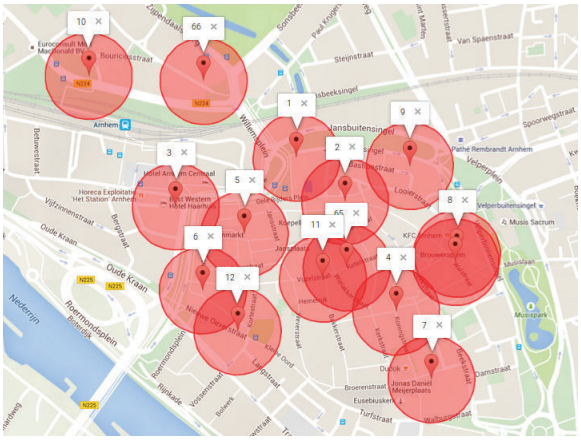

Fig. 2. Wi-Fi Hotspots Placement in Arnhem

We have evaluated this data using the following *metrics*: the number of devices detected by each hotspot, the number of transitions between hotspots, the time spent by each device in the range of each hotspot and the travel speed of devices. The analysis layer not only produced statistics about the visits and the routes but also offered us insights regarding the data correctness. The outlier detections module permitted us to identify problems in the collection process and to offer suggestions for better future deployments.

The Arnhem data mapped into *two scenarios*: the Festival scenario and the Popular areas scenario. The placement of hotspots and their availability did not offer a fine grained view of the people movements during the week, so for the non-festival data we focused on pedestrian densities more than on routes.

The raw dataset contained 2,428,426 records of device connections, containing a total of 33,934 unique devices, between 2014-09-27 18:31:36 and 2014-10-12 21:38:00. Selecting records for the two festival days, the number of detections dropped to 2,375,956 with 32,713 unique device ids. This numbers dropped after the *outlier analysis*, in which we eliminated devices with too many detections (based on their percentage) and with too few detections (less than five). We identified transitions performed in a very short interval (a few seconds) between non-overlapping hotspots and eliminated them from our movement analysis. This step helped us find a problem in the infrastructure that collected data: some hotspots had an incorrect implementation of the hashing function that made it possible for two devices detected by different hotspots to have the same hashed ID. Elimination of these records resulted in fewer correct routes, but it did not affect the density analysis.

The *validation* we performed on these datasets involved identification of the day and night pattern and weekends comparisons. We evaluated the entire period and it confirms the day and night pattern, a snippet for a few days being shown in Figure 3. We can observe how the number of distinct devices starts increasing in the morning, achieving a peak at 1-2 PM (5000 unique devices) and after this point it starts decreasing to a night level. On Saturday the number of detected devices drops by almost 50 percent, meaning that during the weekend people do not visit the down town area so much as in the week days. By comparing the "normal" Saturdays with the one from the festival, we noticed a double up of the number of detected devices, meaning that the city-scale festival indeed crowded the city by increasing the densities of the areas where the Wi-Fi hotspots were placed.

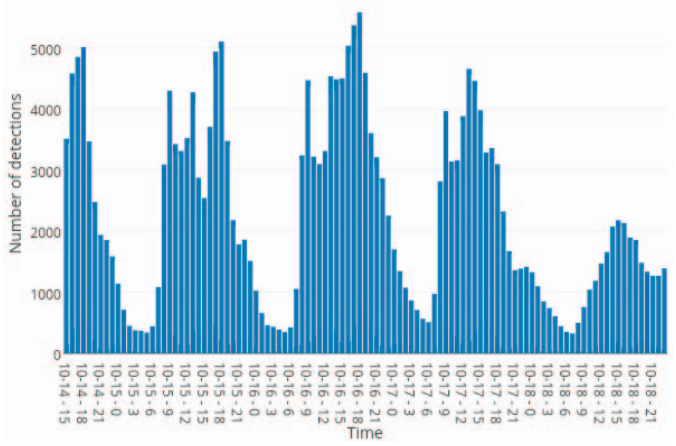

Fig. 3. Total number of detections between Tuesday, 14 Oct 15:00 and Saturday, 18 Oct 21:00

We considered that an intuitive way to show visualizations for our densities and even flows is by plotting them over a map overlay. For the dynamics of the pedestrians and the flow between different areas of the city, we plotted heat maps for the areas covered by our Wi-Fi hotspots. Figure 4 shows one example of a heatmap; we generate a dynamic heatmap for a user-defined interval, and by viewing it dynamically, in a video-like way, one can observe the crowd flow from one hotspot to another. This visualization component makes the transition between density and movement analysis. In order to create a chronological link between consecutive heatmap snapshots, the density values are normalized using the median value of all densities. The user interface shows the heat map dynamically, in an animation covering the whole festival interval and we can better observe the changes and transitions between various areas of the city.

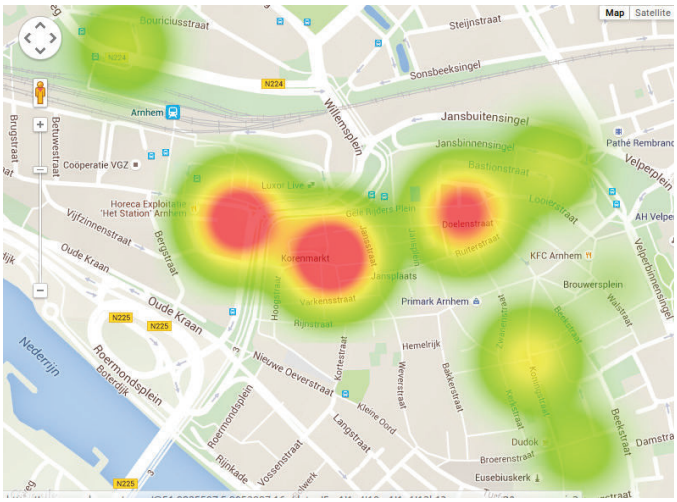

Fig. 4. Density heat map

Our system also allows plot-based comparisons between hotspots, the user selecting the hotspots (any number) and the time interval, such as in Figure 5.

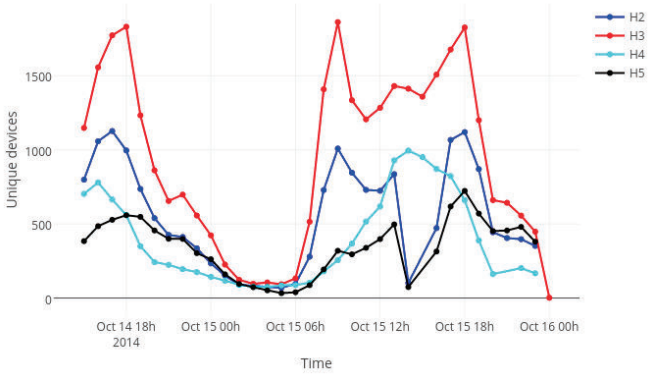

Fig. 5. Hotspots densities comparison during several days for four hotpots

The mobility analysis showed that from the total of 33,934 of unique devices, only 11,567 of them were detected by two or more Wi-Fi hotspots. Further more, this number drops again, after the framework filters the false positive transitions: a device was detected by two different hotspots, but the time interval between the consecutive detections is bigger than the walking time needed to get from one area to another. The remaining 789 devices recorded a total of 2,393 transitions. By computing the transitions between hotspots for both directions we can confirm some of the parts of the festival route proposed by its organizers.

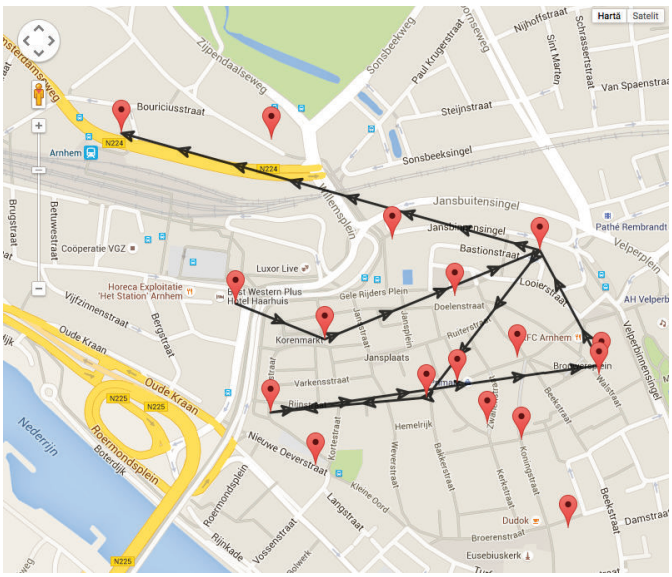

Fig. 6. Individual route plot example

Since the dataset was collected in The Netherlands, a country well known for its high usage of bicycles on a daily basis [5] it made sense to analyze the types of participants in the crowd: the ones who walk and the ones who cycle. From the total of 789 recorded devices with 2,393 transitions, the framework has classified 431 of them being pedestrians and 284 being bicycle users, based on their transitions speeds. The remaining 74 devices had inconclusive transition speeds and they could not be classified. The average speed for the pedestrian transitions for this dataset is 1.25 m/s, and the average speed of people using bicycles is 4.92 m/s.

We concentrated our route analysis on the festival dataset, since the organizers were also interested in checking if their proposed attractions route was used or not. By computing routes followed by each device, we can confirm some of the parts of the proposed festival route. For routes and flow between hotpots we have several visualization techniques, one of them presented in figure 6 where we plotted the individual route of a device. The framework allows us to customize such a view with timestamp labels or other devices.

B. *Discussion*

The accuracy of crowd mobility in urban spaces analysis is strongly dependent on the quality and quantity of the recorded dataset. The quantity refers to the spatial and temporal coverage of the recorded detections. The more hotspots are placed around the areas of the city, the more information is collected and then analyzed. The results of the gathered detections help city planners and authorities to better assess the dynamics of crowded areas and present density dispersion solutions. Moreover, a big spatial granularity makes the localization of a device more precisely. Considering the Ping Pong effect (a device is in the area of two Wi-Fi hotspots and connects from one to another, depending the received signal strength) and the strength of the Wi-Fi signal, a better localization precision is achieved. The temporal quantity of the data takes into account the collection of data for several days and weeks, so that cyclic patterns emerge and also a statistical model about crowd density and movement can be established. This model could be used to prevent dangerous events when densities rise above the normal baseline. Having to work on a real dataset brings some challenges concerning the quality of entries, something that affects the outcome of the analysis. In the datasets we studied, the challenges were represented by inconsistencies and incorrect values in the entries present in the database. These kind of entries were removed from the actual analysis or were repaired and adjusted to some degree. By classifying the source of these challenges, we divided them into two categories: collection system design flaws and analysis logic errors. The first category dealt with errors in hashing function and assignment of timestamps. The Arnhem dataset assigned timestamps when inserting the detection into the central database, this mechanism being prone to lag when the density of people increases. A better approach is for hotpots to use synchronized clocks and include the timestamp in the detection information they send to the server. The transition time computation is also challenging when we know only the hotspot locations and their detection areas. To better assess the speeds we would need their RSSI values and their positioning should provide good street coverage. For example, in the Arnhem dataset some hotspots were far from each other and people could have taken various routes between them, making it hard for us to determine their speed. Another challenge in dealing with such data is what validation strategies to employ, since we do not have access to other statistics to compare with.

V. Conclusions and Future Work

In this paper we presented several methods for analyzing the evolution of crowd densities and the mobility of pedestrians

in urban environments. We relied on detections of mobile devices such as smartphones and tablets recorded by a distributed collection system formed by several wide range Wi-Fi routers. In such a context, visualizations play an important role and we looked for the most appropriate ones for each type of analysis. We incorporated these techniques in a layered system interfaces that offers support for the whole process, from importing a dataset and cleaning it, to visualizing densities and movement around the monitored area. We applied this solution on data collected during an experiment that lasted several weeks in Arnhem, The Netherlands and obtained insights about the crowd densities, classified the type of transportation and identified popular routes and flows inside the city center.

Using the system presented in this paper, we will further analyze other datasets and focus on developing density and movement predictions, which were not possible on the current datasets due to lack of data. Using larger and more fine grained sets (that also offer signal strength readings), we can train neural networks to understand the evolution of densities during a day or a week, and to further generate the possible values for the next hours or days. The same principles can be applied to transitions, by predicting movement of crowds based on their traveling history. Another aspect we want to improve is the accuracy of the analysis by including in our system data from other sources, since we had access to only Wi-Fi hotpots traces, which do not cover people with no Wi-Fi enabled devices.

REFERENCES

[1] A week on Foursquare analysis. http://flowingcity.com/visualization/a-week-on-foursquare/. [last accessed 26-Jul-2015].

[2] Cities growth. http://tinyurl.com/nmm37x5. [last accessed 26-Jul-2015].

[3] EWiDS project website. http://www.distributed-systems.net/index.php?id=extreme-wireless-distributed-systems. [last accessed 26-Jul-2015].

[4] Mobile market growth. http://a16z.com/2015/06/19/mobile-it-changes-everything/. [last accessed 26-Jul-2015].

[5] Statistics for daily bicycle usage in EU. http://www.ecf.com/press-corner/cycling-facts-and-figures/. [last accessed 26-Jul-2015].

[6] U. Blanke, G. Troster, T. Franke, and P. Lukowicz. "Capturing crowd dynamics at large scale events using participatory gps-localization". In 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP).

[7] C. Chilipirea and A.-C. P. Petre. "The GoFlow Project - Robust detection of WiFi-enabled smartphones - Monitoring urban dynamics based on wireless data", 2014.

[8] D. Helbing, A. Johansson, and H. Z. Al-Abideen. "Dynamics of crowd disasters: An empirical study". Physical review E, 75(4):046109, 2007.

[9] A. Johansson, D. Helbing, H. Z. Al-Abideen, and S. Al-Bosta. "From crowd dynamics to crowd safety: a video-based analysis". Advances in Complex Systems, 11(04), 2008.

[10] T. Liebig. "Pedestrian Mobility Mining with Movement Patterns". PhD thesis, Universit¨ats-und Landesbibliothek Bonn, 2013.

[11] T. Nishimura, T. Higuchi, H. Yamaguchi, and T. Higashino. "Detecting smoothness of pedestrian flows by participatory sensing with mobile phones". In Proceedings of the 2014 ACM International Symposium on Wearable Computers. ACM, 2014.

[12] L. Schauer, M. Werner, and P. Marcus. "Estimating crowd densities and pedestrian flows using Wi-Fi and Bluetooth." In Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. ICST, 2014.

[13] A. Stopczynski, J. E. Larsen, S. Lehmann, L. Dynowski, and M. Fuentes. "Participatory bluetooth sensing: A method for acquiring spatio-temporal data about participant mobility and interactions at large scale events". In 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops).

[14] M. Wirz, T. Franke, D. Roggen, E. Mitleton-Kelly, P. Lukowicz, and G. Troster. "Probing crowd density through smartphones in city-scale mass gatherings." EPJ Data Science, 2(1), 2013