Analysis Methods for Extracting Knowledge from Large-Scale WiFi Monitoring to Inform Building Facility Planning

Antonio J. Ruiz-Ruiz
Department of Computer Engineering
University of Murcia, Spain
Email: antonioruiz@um.es

Henrik Blunck, Thor S. Prentow, Allan Stisen and Mikkel B. Kjærgaard
Department of Computer Science
Aarhus University, Denmark
Email: blunck,prentow,lans,mikkelbk@cs.au.dk

***Abstract*—The optimization of logistics in large building complexes with many resources, such as hospitals, require realistic facility management and planning. Current planning practices rely foremost on manual observations or coarse unverified assumptions and therefore do not properly scale or provide realistic data to inform facility planning. In this paper, we propose analysis methods to extract knowledge from large sets of network collected WiFi traces to better inform facility management and planning in large building complexes. The analysis methods, which build on a rich set of temporal and spatial features, include methods for noise removal, e.g., labeling of beyond building-perimeter devices, and methods for quantification of area densities and flows, e.g., building enter and exit events, and for classifying the behavior of people, e.g., into user roles such as visitor, hospitalized or employee. Spatio-temporal visualization tools built on top of these methods enable planners to inspect and explore extracted information to inform facility-planning activities. To evaluate the methods, we present results for a large hospital complex covering more than 10 hectares. The evaluation is based on WiFi traces collected in the hospital's WiFi infrastructure over two weeks observing around 18000 different devices recording more than a billion individual WiFi measurements. For the presented analysis methods we present quantitative performance results, e.g., demonstrating over 95% accuracy for correct noise removal of beyond building perimeter devices. We furthermore present detailed statistics from our analysis regarding people's presence, movement and roles, and example types of visualizations that both highlight their potential as inspection tools for planners and provide interesting insights into the test-bed hospital.**

I. Introduction

Healthcare administrators are constantly under pressure to reform the healthcare system organization by planning activities to better utilize available resources to minimize cost but at the same time offer a high quality healthcare service [1], [2]. The design and maintenance of a cost-effective and high quality healthcare system is an ongoing high-priority challenge for most governments around the world. A crucial part of this challenge is the difficulties inherent in planning hospital activities—as these require an accurate knowledge of the hospital environment, of the availability of resources (both materials and personnel), of knowledge about flows of personnel and patients, and usage of services and facilities. One example where better planning can help optimizing healthcare services are removal of inefficiencies in patient flows, e.g., of patient misplacement or of patient late arrivals which result in surgery cancellations [3].

Today, only statistics from patient records are generally available to hospital facility planners [2], e.g. number of ambulant treatments and hospitalizations. Other existing approaches [4], [5] have tried to address the lack of knowledge using a modeling approach. These approaches focus on length of stay and flow of patients between departments to provide models reflecting the complex, variable, dynamic and multidimensional nature of hospital systems. However, in [6] the authors demonstrate that such model-based calculations typically do not provide the appropriate information needed to obtain reliable results—since the models do not take into account all variables influencing the continuous operations at a hospital. Examples of such variables include: i) amount and spatio-temporal distribution and flow of visitors—influencing the planning of offered facilities such as seating areas, parking spaces, and toilets; ii) precise up-to-date information about people within the building complex such as their role as patients, visitors, and staff.

Nowadays, widespread user devices such as smartphones, tablets and in the future also smart watches, emit WiFi signals on a frequent but irregular basis [7]. Moreover, the already available wireless infrastructures in large building complexes, like hospitals, enable the collecting of large data sets of WiFi measurements that can be used not only to analyze the network's performance and usage, as proposed in earlier work among others [8], [9], [10], but potentially also the density and flow of people within the building. Compared to earlier approaches based on Bluetooth, in urban [11] or indoor settings [12], or based on video in indoor settings [13], the use of WiFi comes with lower setup costs, due to the existing deployment, for monitoring complete large-scale building complexes. However, analysis methods are missing that allow to extract information, relevant for planning, from collected large-scale WiFi data sets.

In this paper, we propose analysis methods to extract knowledge from large sets of WiFi traces to better inform facility planning in large building complexes. The analysis methods build on a rich set of temporal and spatial features extracted from the WiFi traces. The analysis methods include methods for i) noise removal, ii) quantification of people densities and

flows per area of interest, and for iii) classifying the behavioral roles of people. To remove noise we propose methods to clean data, filtering out, e.g., device traces that are close to the perimeter of the building complex but not within it. We do so by labeling these devices as beyond building-perimeter devices using machine learning-based classification with a novel set of features calculated from raw WiFi signal data. For estimating people densities and flows in areas we propose heuristics to filter streams of calculated device positions—assessing, among others, the number of enter and exit events. To classify user roles we propose a method based on machine learning-based classification on our rich set of spatial and temporal features, trained with weakly labeled data. Furthermore, we present spatio-temporal visualization tools built on top of the described methods for enabling planners to inspect and explore extracted information to inform facility planning activities.

To evaluate the proposed methods, we present results for a large hospital complex covering more than ten hectares in which we have collected WiFi traces over two weeks observing around 18000 different devices recording more than one billion individual WiFi measurements. Moreover as background information we also present detailed statistics of the observed devices, e.g., type of devices and the frequency of observations. We present quantitative results for the analysis methods, e.g., for noise removal of beyond building perimeter devices where results demonstrate over 95% accuracy for correct removal. For the quantification of area densities and flows we present comparisons with manually recorded flows. For the classification of user behavior we present results showing a high degree of correlation with statistics provided by the test-based hospitals regarding visitors, staff and hospitalized people. Additionally, we present example visualizations such as heat and flow maps that both highlight the visualizations' potential as inspection tools for planners and provide interesting insights into the hospital's workings.

The presented methods can be generalized and thus applied not only to hospital settings but enable facility analysis also in other types of large building complexes such as industrial facilities, shopping malls or public buildings in general. The proposed methods can be also used to analyze the spatio-temporal distribution of people to offer better planing services and facilities, e.g., seating areas, parking spaces, toilets, and their maintenance, e.g., for cost-efficient scheduling of cleaning personnel at times of low load on the respective facilities.

II. Related work

Existing work utilizing measurements from wireless networks [8] focused on analyzing the networks' performance and usage. The analysis was based on aggregating the data into various forms of graphs and statistical summaries; for instance, to obtain statistics about the number of devices that made use of the network, which applications the network was used for, and the mobility of the users. The main aim of these studies was to improve the design, modeling and management of wireless networks in regards to, e.g., improved protocol designs or better adaptability for areas where APs exhibit a lot of network traffic. Such studies have been performed both in an university campus settings [8], corporate settings [9] and urban settings [10]. For a campus setting Calabrese et al. [14] proposed methods to explore overall user behavior for buildings on the campus but did not relate it to the within-building movements.

Another line of work has utilized data collected from people's own devices instead of using data from wireless networks. Such work has analyzed different aspects of people's behavior and of the places they visit. Chon et al. presented a system for categorizing places from mobile device data [15]. Vu et al. [16] presented a framework for constructing predictive models of people's movement. Focusing on sensing of the collective behavior of crowds, different methods have been proposed, e.g., to estimate properties regarding flocking, followers and density. Kjærgaard et al. [17], [18], [19] propose methods for flock detection and follower detection based on mobile sensing data. Neil et al. [11] consider methods for counting people in an urban setting using Bluetooth scanning. Other approaches focus on traffic analysis, including Musa et al. [7], and study vehicle tracking based on passive WiFi transmissions. The above study demonstrated that tracking unmodified devices using WiFi monitoring is feasible in outdoor settings but it did not consider indoor settings or facility planning. In contrast to previous work in this paper we propose analysis methods utilizing data from WiFi networks in large building complexes. These methods are designed to extract knowledge from such data to inform facility planning.

III. Hospital Testbed

During the process of developing the proposed analysis methods we have collaborated with staff from the planning and IT departments at the Aarhus University Hospital. In discussions the staff told that their current practices for planning are mainly based on statistics from patient records and coarse estimates which is common according to existing research studies [2]. Furthermore, they were very interested in new means of obtaining and using more realistic information for their planning activities.

In collaboration with the hospital a data set of Wi-Fi measurements was collected through-out the hospital complex, see Figure 1 for an overview of the complex. The hospital features 22 different buildings with up to 3 stories, covering an area of more than 10 hectares. The entire hospital relies on a wireless network infrastructure that covers all of its buildings, with the exception of those areas reserved to surgery rooms, where, due to safety reasons, electromagnetic radiation is restricted. The total amount of access points (APs) available in the hospital is 798, with most of them (around 95%) being

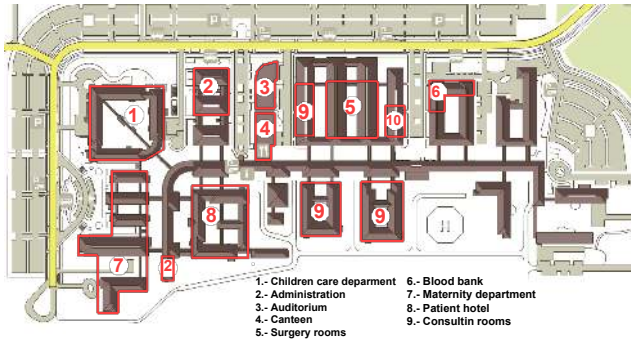


Fig. 1. Aarhus University Hospital - Skejby complex.

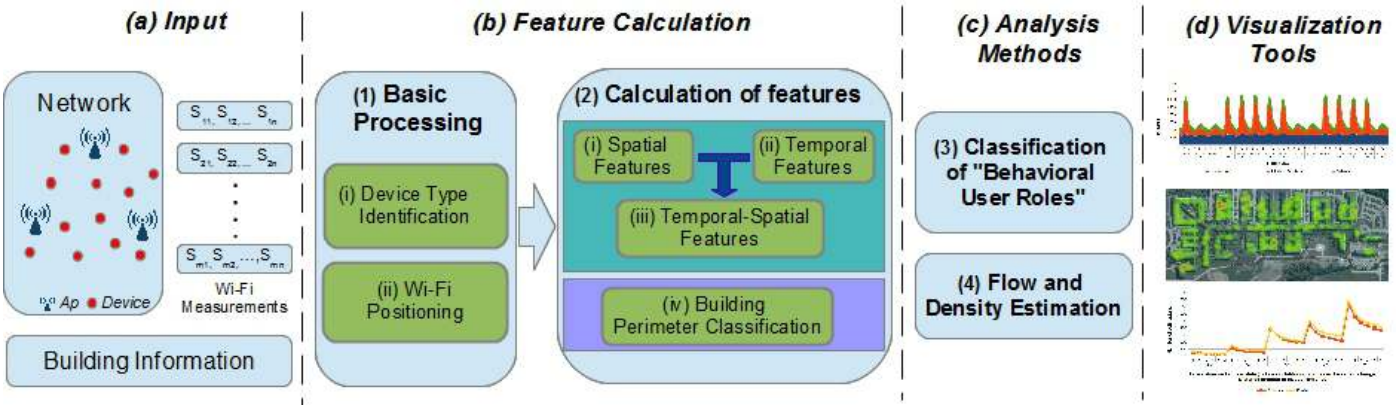


Fig. 2. Overview of the steps involved in data processing feature extraction, analysis methods and visualization.

Trapeze and Juniper devices. The network provides several virtual networks including a guest network open to the general public. The system architecture used for data collection is network-based, i.e., WiFi measurements are collected by the APs on all WiFi channels and forwarded to a central server which stores them to a database. Our data collection was carried out for 15 days using all available APs, collecting in total more than a billion of WiFi measurements from around 18000 different devices.

One important aspect in large-scale mining studies is that some of the extracted features (e.g. user position) are privacy sensitive—especially when working in hospital environments, since personal health information must be protected in regards to identification of individuals. Regarding this concern, we emphasize that we only collected network scan frames, and used an anonymization procedure during data acquisition that ensures a high level of privacy protection. Following the same approach as utilized for the Nokia data challenge [20], MAC addresses were encrypted by hashing after concatenating them with a secret key. This ensures that the collected tracking information can not be re-associated with a specific device.

IV. Overview of Analysis Methods

Figure 2 illustrates how the proposed analysis methods build upon each other and together enable a tool chain to extract knowledge for facility planning and provide associated visualizations. The data used as input (a) are provided from two different sources: WiFi measurements from a large-scale wireless network, and a geometric model of the perimeter of the building complex. The feature calculation phase (b) is divided into two steps consisting of: (1) basic processing where the type of the device is identified $(1.i)$ and the raw WiFi measurements are converted to positions using existing WiFi positioning algorithms $(1.ii)$; and (2) calculation of a rich set of spatial $(2.i)$, temporal $(2.ii)$ but also spatio-temporal $(2.iii)$ features to enable the analysis methods. In addition, our novel technique to deduce from raw signal measurements whether a device is within the building perimeter is applied $(2.iv)$. The analysis methods (c) extract relevant information from the calculated features. The proposed heuristic based-method for quantifying densities and flow (4) is applied to estimate the usage of entries and exists. Furthermore, for classification of behavioral user roles (3) we apply machine-learning, and in the hospital setting we apply it to classify devices according to a device owner's role as either a short-term visitor, long-term visitor, employee or hospitalized person.

The visualization tools (d) provide intuitive and interactive access to the information extracted in order to facilitate assessment and planning regarding facilities and services in the building complex. The visualization tools show different outputs provided along the entire process as heat-maps, flow-maps, graphs and tables and thereby provide an important set of information that reflects different aspects of the utilization of the buildings, and of associated facilities and services.

V. Feature Calculation

This section covers the proposed rich set of features calculated to enable the mentioned analysis methods. Furthermore, to argue for the feasibility of using large-scale WiFi traces for facility planning we provide illustrating examples of the feature data calculated from the hospital data set.

A. Device Type Identification

Network-based WiFi monitoring effectively collects measurements for all signal emitting WiFI devices—both for mobile as well as for infrastructure devices. When analyzing densities and flows of people we are only interested in mobile devices; we therefore aim to filter out the infrastructure devices. To this end, we apply a filter which exploits that the devices' MAC addresses encode the manufacturer in the initial three octets. This allows to distinguish manufacturers who mainly produce mobile devices (Apple, HTC, Samsung, Sony, Nokia, Huawei, Murata, LGE and RIM) from those producing mainly infrastructure devices (Cisco, Trapeze and Juniper). Figure 3 shows the distribution of devices we observed, grouped into infrastructure, mobile, and devices of undetermined type: The amount of infrastructure devices remains stable over time, while the amount of mobile devices regularly increases during day-time, and decreases over weekends and holidays (here: Friday the 26$^{\text{th}}$).

Figure 4 compares the observed disribution (blue) of devices, subclassified according to operating system, with statistics about smartphone platforms in Denmark during 2012[1] (red). The obtained distribution of the different platforms

[1]Statistics provided by Google's Our Mobile Planet service, http://www.ourmobileplanet.com

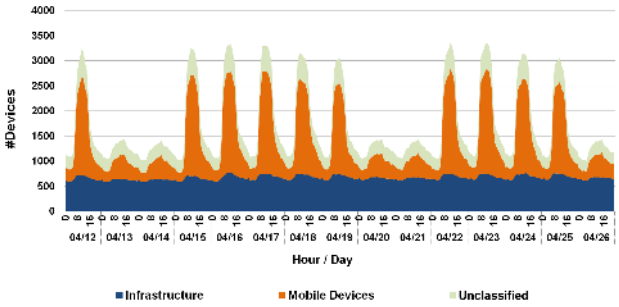


Fig. 3. Amount of devices, grouped by device type, observed over 2 weeks.

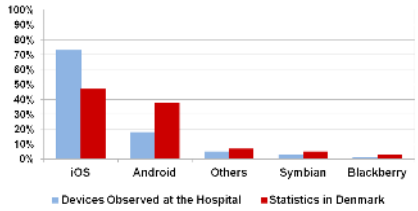


Fig. 4. OS distribution among mobile devices observed at a hospital.

fits the statistics as obtained for the general population. The apparent differences, prominently the larger percentage of observed iOS devices is in part due to that the observed amounts of mobile devices also includes tablets and pods, since the MAC address analysis does not allow to distinguish them from smartphones. This concurs with that 33% of danish households have a tablet and most of these are iOS devices.[1]

B. Large-scale WiFi Positioning

To estimate the position of the observed devices we use a WiFi positioning module. Since we do network-based measurement collection we will only be able to position devices when they scan for networks. Musa et al. [7] provide statistics and observations of the scanning behavior of different mobile devices, e.g., most devices scan when the screen is turned on or when they aim to transmit data. In the collected data, the median and average time between a mobile device's scans are 58 and 196 seconds, resp. Whenever an AP observes a scan it sends to a central machine a measurement message which contains: the id of the AP, the MAC address of the device, the received signal strength (RSS) in dBm, and a timestamp. The main advantage of using this network-based measuring approach is that every device providing WiFi connectivity can be monitored, independently of its platform and installed software, thus reducing the system deployment time and cost and not requiring the user to install specific software [21].

At the central machine MAC addresses are encrypted and position estimates are computed from the RSS measurements using the centroid lateration algorithm as described in [22]. For this computation the algorithm only requires the location of the APs. Using these, the algorithm estimates the position of a device to be the weighted geometric average of the locations of the receiving APs, using as weights the received signal strengths for each AP. The estimate is then snapped to the location of the nearest AP, in order to enforce that reported positions are inside the buildings. Using this approach the position estimates were evaluated to have a mean accuracy of $15m$ on traces collected through-out the buildings. While other methods may provide more accurate estimates, such as fingerprinting based methods [23], they have additional requirements such as collection of fingerprints or the availability of digital building models. Reliable fingerprint collection (and keeping it up-to-date over time, facing also building- and WiFi-infrastructure changes) at a hospital with more than six thousand rooms spread over ten hectares was deemed unfeasible [21]; furthermore, a complete digital building model, suitable for fingerprinting, of the hospital was not available.

C. Classification of Beyond Building Perimeter

Discriminating whether a device is inside or outside one of the complex's buildings is a difficult task as such complexes often have court yards and passages between buildings. Previous work has considered this problem using GPS signals [24] and other sensor modalities [25]. However, given only WiFi measurements these solutions do not apply, and the WiFi positioning literature has also not yet addressed the problem.

In general, when being located outside but close to a building, the WiFi signals emitted from a device can be observed by the APs within the building; a positioning module as described above would therefore end up placing the observed device inside the building. These situations generate erroneous cases in which the device could receive certain information, e.g., from an indoor navigation application or advertising from a specific shop, when it is still out of the buildings that offer these services. In the chosen scenario such errors may impair our analysis methods, e.g., for detecting the time of entry in a building. Moreover, distinguishing outdoor from indoor positions may allow us to filter out those devices that never enter the building and therefore should not be taken into account in statistics of people utilizing the building facilities. The outdoor detection method we present here is based on several features which have been extracted from the set of RSS measurements at a given timestamp and the resulting estimated positions: (i) The signal strength difference between the strongest AP and the weakest AP observed; (ii) Average signal strength of the k-strongest APs received; (iii) Averaged distance between the device's estimated position and the position of the k-strongest APs received; (iv) Average distance among all the received APs; (v) Percentage of *perimeter APs* observed: for this we define a perimeter area utilizing the building complex's layout data, and we classify a AP as either a perimeter AP or interior AP according to if it is within or out of the perimeter area. In Figure 5 the perimeter area is highlighted as it is defined for the hospital; note, that this area includes only the part of the perimeter that is physical accessible from public streets. Given this labeling we can calculate the percentage of perimeter APs for each observation. For features ii) and iii) we empirically tested several values for k and we finally chose a value of $k = 3$. We are conscious that the features listed above may need to be adjusted in order to use the classifier at other building complexes according to their wireless network infrastructures. For instance, For high-rise buildings also the floor level detected by the positioning system can provide valuable input to the classifying procedure. Furthermore, our analysis detailed below revealed that among the listed features the ones having the strongest benefit for the intended classification are features iv) and v); these features are largely independent from specific device's hardware characteristics (e.g. from absolute RSSI value computations), and thus can cope well with device heterogeneity [26].

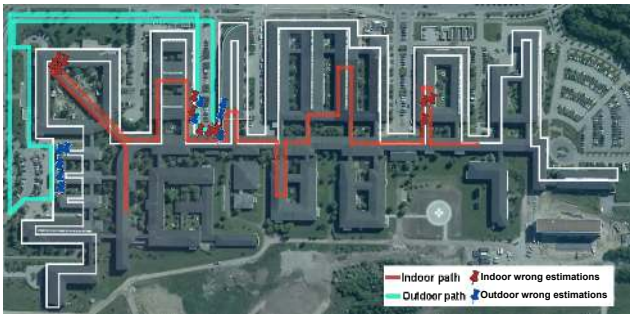


Fig. 5. Geometry that defines the perimeter area of the building complex. Indoor and outdoor paths and examples of wrong estimation cases.

The classifier was implemented using the JAVA API provided by Weka [27]. For initial testing and refining we collected data during two 15 minutes indoor and outdoor walks through the hospital and its surroundings, see Figure 5, using four different devices (a Samsung Galaxy Tab and three smartphones, one Jiayu G2-Plus and two Google Nexus One), obtaining around 1,200 data points (54% inside, and 46% outside buildings). We evaluated classification accuracy using four different classifiers: Random Forest, J48, BayesNet and Decision Tables. Previous to the classification, we normalized the training dataset by applying a resampling filter, generating a uniformly distributed random subsample of the dataset. Table I shows the classification results obtained using ten-fold cross-validation. The results show a high accuracy of over 90% for all used classifiers, which indicates that the extracted features are indeed reliable in differentiating between indoor and outdoor positions. We selected the most robustly performing learner, the Random Forest algorithm, to be used in posterior experiments. As one can observe in Figure 5, most of the incorrect classifications (labeled by colored pushpins) in both test cases, indoors and outdoors, happen near main entrances or in areas where the number of APs detected is relatively low (which is the case in one of the building corners). Although entrances are a crucial challenge when distinguishing inside and outside positions, the problem can be alleviated since people are not constantly leaving and entering a building in short order: i.e, when they are inside or outside the building, they usually remain so for a sufficiently long time to produce several position estimates. This in turn allows us to optimize the robustness of the estimations, c.f. Section VI-A.

	J48	Random Forest	BayesNet	DecisionTable
Correct Inside	535	551	514	527
Incorrect Inside	22	6	43	30
Correct Outside	605	612	563	591
Incorrect Outside	21	14	63	35
Accuracy	96.3%	98.3%	91%	94.5%

TABLE I. CONFUSION MATRIX FOR INSIDE AND OUTSIDE CLASSIFICATION.

D. Calculation of Features

A crucial task for the goal followed in this paper and when dealing with large sets of unlabeled data is the design of features for extracting vital information on which further analysis can build. In the following we list features central to this task, differentiating them into three categories:temporal, spatial and spatio-temporal.

Temporal features capture aspects concerning the times a device is located within the building complex.

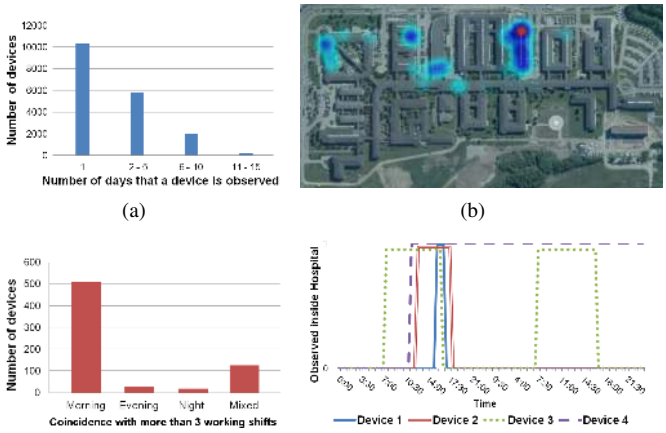


Fig. 6. (a) Number of unique devices grouped according to the number of days they were observed. (b)Areas where a device spent most time stationary. (c)Statistics about working shifts.(d) Time of detection inside the hospital of four different devices representing the four expected behaviors.

Number of days detected (T1) indicates the number of days we observe a specific device, as shown in Figure 6(a). Within the chosen use-scenario the feature helps distinguishing between devices that belong to employees and those that belong to visitors, since the duration of observations should be clearly different in those cases.

Hours per day (T2) spent inside the building complex. In general terms, employees' smartphones remain visible within the building more hours per day than those of short-term visitors, but less than those of hospitalized persons. Such differences can be observed in Figure 6(d), where *Device 1* is typical for a short-term visitor whereas *Device 4* is typical for a hospitalized person.

Daytime (T3) indicates the times of day each device is observed. We distinguish between: during day-time(e.g. 7am to 11:59pm), night-time (e.g. 9pm to 6:59pm) and during both. As shown in Figure 6(d), devices of hospitalized persons (*Device 4*) are usually observed at any time, whereas visitors are mainly observed during daytime.

Working shifts (T4) help us to discriminate what devices belong to employees or other people that have a fixed timetable. Since the ranges of working hours can vary from one environment to another, we have taken into account the hospital working shift schedule (from 7am to 3pm, from 3pm to 11pm and from 11pm to 7am). Figure 6(c) shows the number of devices whose duration inside the hospital correlates with a shift time on at least three days. Those devices would clearly belong to employees.

Spatial features capture aspects of the locations of people (respectively their devices) within the building complex.

Restricted areas (S1) indicates that a device resides within hospital areas that are restricted to certain kinds of people; for example, surgery rooms and laboratories. The areas accessible only to employees are indicated in Figure 1. Moreover, in the particular hospital most parts of the basement floors are only accessible to employees. This last restriction, *times observed in basement* is one of the features that will be used in the posterior processing.

Frequent places(S2) determines the set of areas were a device is frequently observed. This information allows, e.g., to infer ambulant treatment types or job roles.

Beyond Building Perimeter Classification (S3) has been described in V-C and is listed here for completeness.

Spatio-temporal features consider both spatial and temporal aspects of a device's movement within the building complex.

Motion speed (TS1) depicts average speed of a device. The feature's accuracy depends on realised positioning accuracy as well as frequency. We estimate speed based on the distance covered over time. Though this does not provide a highly accurate speed estimation, it serves well to differentiate motion status (still vs. moving) of devices. Earlier work [28] has proposed a more accurate method for still vs. motion detection using raw signal measurements, however, we did not apply this method because it requires frequent measurements often not satisfied in our data set.

Time stationary (TS2) reflects whether a device has been stationary for a longer period of time—which we define here as being located for more than T minutes within r meters of any single place. For choosing r we suggest taking into account the average distance among APs.

Places where stationary (TS3) determines, relating to the feature S2, the different locations where a device has been stationary, e.g., in a waiting, patient or meeting room. Figure 6(b) shows a building map indicating the places visited by a device during one day (with the color scale indicating total stationary time at the respective locations).

The presented features form the basis for the analysis methods presented in Section VI. Furthermore, the graphical presentation of the collected data set for the described features illustrate and highlight their utilization, revealing e.g., that ca. 2000-3000 mobile devices were observed per day (Figure 3), and that a large fraction of these were observed only on one day (Figure 6(a)). These numbers support that our measurement approach provides rich data for a significant number of devices. Compared to previous wireless network studies in campus or company settings [8], [9], [10], the large percentage of one-day-only visitors differentiates this data set from what has been observed in the above studies where the sets of perceived devices per day were highly correlated across days. This also highlights that hospital environments are different and thus relevant use-scenarios to consider in future work in wireless network analysis and related fields.

VI. Analysis Methods

In the following, we describe how to utilize the features extracted from WiFi measurements for further analysis methods for informing and supporting decisions within facility planning. In particular, we propose methods in two areas, briefly introduced already in Section IV.

A. Density and Flow Estimation

The density of people in a specific area or the flow of people through a given area or across a given line or other borders are fundamental types of information within planning in both indoor or urban settings [11]. To obtain such information, we propose to apply a number of heuristics using the features introduced in Section V. In the following, we will consider the specific case of quantifying the flow through entrances as people enter and leave the hospital. Such information enable the deduction of, e.g., the most used entrances to a building complex, which helps to decide e.g., where to install information boards or vending machines (since these would be the most busy areas), or to determine the flow-wise most appropriate entrances for emergency cases (i.e.,less crowded ones), or to determine where to build additional parking places (i.e., in those areas by which people usually enter into the hospital), or to design evacuation plans (for individual daytimes or weekdays, or even dynamically, according to current crowd conditions, among others).

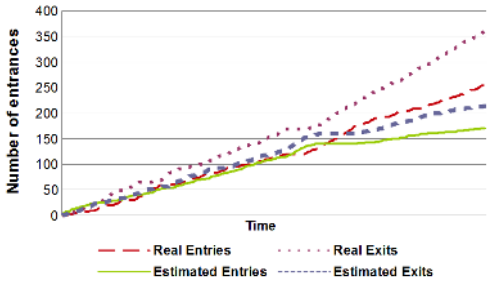


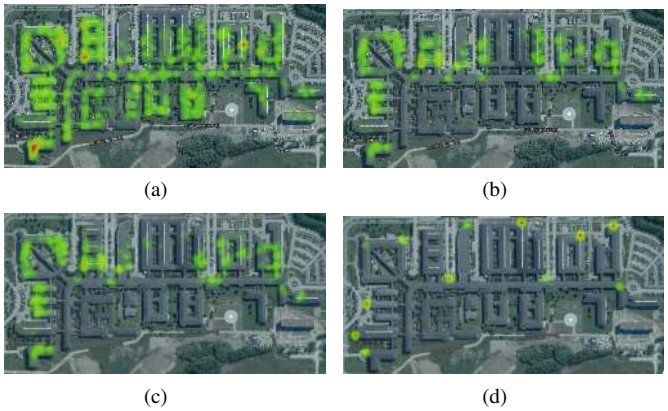
Fig. 7. Entry and exit events over time for $S = 30s$ and $R = 40m$.

To estimate the flow through entrances we propose a method building on the beyond building perimeter classification from Section V-C. Having calculated the beyond building perimeter feature value, once we detect a change in the device's in/outdoor state, we record its timestamp. To avoid erroneous rapid state changes provoked by signal variability, the method waits for the new state to remain stable for at least S seconds before it registers a new entry or exit event. We assign the event to the closest entrance among a list of entrances previously defined. To avoid false positive cases we record the event only in case the distance between the closest entrance and the estimated position is below a threshold R.

To evaluate the method's accuracy, we have carried out several empirical tests using different configurations for the threshold parameters S and R which define whether an entry/exit event should be recorded. Figure 7 depicts the number of entry and exit events that have been estimated at the hospital's main entrance over a period of 6 hours. During this time, a person manually counted the number of actual entries (327) and exits (453) at the entrance, obviously with no knowledge about how many of people that were counted also carried a smartphone. We can assume that the ratio of smartphone holders is close to the 59% reported as the estimated percentage of smartphone penetration in 2013 in Denmark[1]. These numbers would correspond to 192 entries and 270 exits of persons with smartphones approximately. Using the configuration values $S = 30$ seconds and $R = 40$ meters we are able to approximate the expected results as shown in Figure 7. Finally, in order to provide visualizations of various obtained results on the complete data set we build, after executing the method, heat-maps as shown in Figure 8.

B. Classification of Behavioral User Roles

In this section we propose a method to classify the behavioral roles of users. For this classification we employ manually constructed as well as machine-learning rules, both of which operate on the spatial, temporal and spatio-temporal features



(a) (b) (c) (d)

Fig. 8. (a) Heat-maps representing all device positions; (b) Only inside to outside movements (leaving the building); (c) Positions of filtered exits; (d) Estimated exits constrained to real exits.

described in Section V. Knowing the behavioral roles of device owners (such as staff, visitor, or patient) is useful, e.g., for compiling and interpreting statistics regarding the utilization of individual facilities or services within a building complex. In particular, for planning and optimizing services and facility utilization it is crucial to know the current utilization by individual user groups such as staff or visitors. E.g., for the chosen environment type, it is important to consider if visitors use entrances and paths through the building designed for employees, e.g., entering parts of the hospitals where visitor traffic should be kept to a minimum, or where they are not allowed to enter. Furthermore, when analyzing the paths used by visitors (resp. staff), it is important to filter out traces from staff (resp. visitors). For the chosen hospital environment, most people can be semantically classified into one of the following four roles: Employees, Hospitalized, Short-Term (ST) Visitors and Long-Term (LT) Visitors. Here, the ST visitor role captures people who pay a hospitalized person shorter visits or who receive simple ambulant treatments, while the LT visitor role captures people receiving more extensive ambulant treatments or people accompanying persons receiving such treatments.

It is not possible to extensively gather manually labeled data for all user roles in order to train machine learning algorithms to differentiate people into these roles, given the special privacy requirements in hospital settings. E.g., contacting people at the entrance and inquiring their role and their phone's MAC address was not considered an ethically viable procedure. Therefore, we instead employed the help of domain experts to derive a set of human-made rules (on basis of our features from Section V) which describe the expected behavior of the individual roles. To this end, first a strict set of rules, shown in Table II, was derived, which only provided a classification when the input data very unambiguously can be classified into one role. The table states these rules—by way of listing both used features as well as the respective feature values expected according to the common behavior patterns for the individual roles. We chose this strict set of rules to classify very conservatively in order to obtain valid and trustworthy separators between roles.

Applying these strict rules allows to classify and assign roles (comparatively confidently) to just 5% of the detected devices. Following a normalization for obtaining a uniform distribution across roles, these classifications for role-assigned devices are then used as training data-set for the machine-learning classification algorithm. We chose the Bayesian Net learner, provided in the Weka data-mining library [27], because it has been found to be reliable for classification purposes when the distribution of attribute values is widely spread; moreover it also adapts well when new training data are dynamically added to the model [29]. Thus, the system can evolve over time, as it is used—allowing the adaption to a wider set of behavioral patterns within each role. For each instance (device) and role, the classifier estimates from an instance's feature values the probability of belonging to each role, respectively. In the first row of Table IV results are given from running the algorithm on the complete data set including the already labeled data.

Feature	Employee	Hospitalized	ST Visitor	LT Visitor
T1	> 5	> 3	1	$2 \mid\mid 3$
T2	$6 <> 9$	> 16	< 0.5	$2 <> 5$
T3	Daytime $\mid\mid$ Nighttime	Daytime & Nighttime	Daytime	Daytime
T4	>5	0	0	0
S1	>50	<10	0	0
TS2	False	True	False	False

TABLE II. CRITERIA FOR STRICT RULE-BASED CLASSIFICATION.

Feature	Employee	Hospitalized	ST Visitor	LT Visitor
T1	$>= 3$	$>= 2$	$= 1$	$>= 2$
T2	$5 <> 10$	> 10	< 1	$1 <> 6$
T3	Daytime $\mid\mid$ Nighttime	Daytime & Nighttime	Daytime	Daytime
T4	>2	0	0	0
S1	>20	<20	< 10	< 10
TS2	False	True	False	False

TABLE III. CRITERIA FOR RELAXING RULE-BASED CLASSIFICATION.

As an alternative to the machine learning algorithm, we have carried out a rule-based classification procedure: We defined, again with domain experts, a more relaxed set of rules: Table III shows the relaxed rules which classify a larger portion of devices into roles, as compared to the rules given in Table II. Note, that the new rules still ensure that the rule-based definitions for roles are not overlapping; i.e., no device can be classified into more than one role. The classification results obtained are given in the second row of Table IV.

Table V summarizes the results obtained with both algorithms and compares them to statistical data. For each role the table lists the number of devices classified into the role (as well as in brackets the percentage among total devices detected). For comparison, we also list for each role the expected number of observed people for a 15 day period, as extracted from annual hospital-provided statistics. First, for hospitalized patients (*) the expected number is obtained by dividing the number of testing days by the average length of stay of 3.5 days, and multiply with the average number of occupied beds, as obtained from OECD Health Data 2009. For day visitors (+), we sum the hospital-provided annual number of treatments (245554), and the product of the number of hospitalized people and the hospital-observed average number of visits per patient and day). To compare to our number of detected visitors, we for the latter aggregate the short-term (ST) and long-term (LT) visitors into one group. In view of these results we can conclude that both the Bayesian Net algorithm

Method	Employee	Hospitalized	ST Visitor	LT Visitor	Unknown
Bayes-Net	1609	359	8290	6837	1241
Relaxed Rules	1206	389	8465	4313	3964

TABLE IV. RESULTS OBTAINED FOR THE DIFFERENT CLASSIFICATION METHODS.

	Bayesian Net	Relaxed Rules	Yearly Statistics	15 Days Statistics
Employee	1609 (46%)	1206 (38%)	$\sim$ 3500	$\sim$ 3500
Hospitalized	359 (20%)	389 (22%)	412 (Beds)	1766 (*)
ST Visitor	8290 (-)	8465 (-)	-	-
LT Visitor	6837 (-)	4313 (-)	-	-
Daily Visitors	15127 (63%)	12778 (53%)	593.901 (+)	24026

TABLE V. OBTAINED RESULTS AND COMPARISON WITH YEARLY STATISTICS. IN BRACKETS, THE PERCENTAGES WITH RESPECT TO THE 15 DAYS STATISTICS ARE GIVEN.

and the rule-based labeling method provide realistic results, despite smaller deviations from statistics for the employee and daily visitor role. That the correlation with statistics is even higher for the role of hospitalized can be explained by that for this role the behavioral pattern is easier to predict and therefore good approximation rules are easy to conceive.

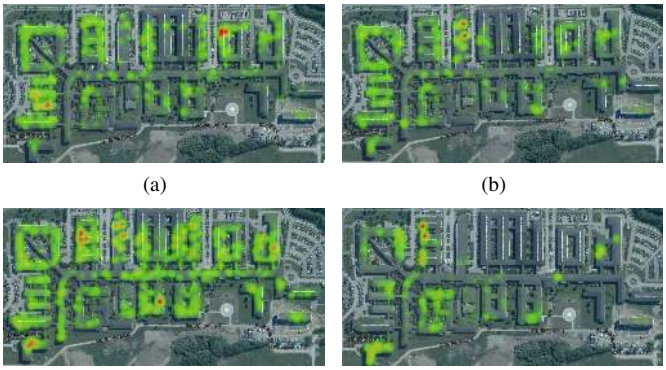
According to statistics[1], the smartphone penetration in Denmark is around 59% in 2013; and around 54% of people (i.e., visitors) use their smartphones while waiting at the doctor's office[1]. The results may be biased by the fact that people not carrying a WiFi enabled device are not accounted for. Nonetheless, our estimations of the number of daily visitors approximate well the statistics provided by the hospital. With respect to hospitalized people, considering that 70% of hospitalized people are over 55 years and 25% are between 35 and 54 years according to hospital information, and taking into account the statistics [1] saying that 27% of people over 55 years and around 71% of people between 35 and 54 makes use of the smartphone at hospitals, we deduce an expected 36% of hospitalized people to make use of smartphones. Note that not all smartphone's users may use them daily during a hospitalization period, what is essential to be classified within this role. That could explain that our method estimates the number of hospitalized people to be 14-16 percentage points lower than statistics suggest, c.f. Table V.

Regarding employees, according to statistics[1] around 84% of people in Denmark use their smartphones at work. In a hospital environment a lower number is expected, since restrictions and tasks for the employees (doctors, nurses and orderlies) do not allow to make use of smartphones and/or the network frequently. Taking into account this fact, together with penetration of smartphones and staff numbers at the hospital, our results provided for employee classification represent realistic numbers.

VII. VISUALIZATION TOOLS

In this section, we present and discuss example outputs of the visualization tools we have built based on the introduced features and analysis methods. The main aim of this section is to illustrate how the latter enable novel in-depth visualizations, and how these can aid those in charge of planning decisions.

To illustrate the use classifications of users' behavioral roles, Figure 9 shows a heat-map for each user role, depicting the positions (while disregarding floor information) for one day and of all devices of that role. The difference among the maps can be clearly observed with regards to the visited areas (refer to Figure 1 for the placement of individual departments). For example, neither ST Visitors nor LT Visitors have been detected near the surgery area, whereas many employees were observed in the vicinity. Moreover, a high number of ST-Visitors come to the hospital for blood donation, or to



(a) (b) (c) (d)

Fig. 9. Heat-maps representing most commonly visited zones by: (a) Visitors Short-Term; (b) Visitors Long-Term; (c) Employees; (d) Hospitalized.

(a) (b)

Fig. 10. Three tests for tracking a device from the used entrance to its stationary destination.

get consultation, or to visit someone at the maternity unit. Employees move around in all the buildings, as expected, whereas hospitalized people are localized in areas that are devoted to suites for patients. Furthermore, animated versions of these plots enable further analysis, e.g., considering the load of different parts of the buildings through-out the day.

To illustrate the methods for flow and density estimation we consider the entry and exit event detection case. The results given in Figure 8(d) visualize the load among the different entries which give rise to several relevant questions, e.g., one of the exits close to the surgery ward has a high load—higher than intended given its location; noteworthy is also that the closest main entrance has a comparatively low load. To further analyze this situation, it is relevant to consider where people using an entrances end up within the building. To enable this analysis we provide additional information about the movement flow observed from the entrance until a device reaches a stationary destination (waiting room, canteen, office, etc.) as described by the "places where stationary" feature (TS3). In Figure 10 we show an example of this type of information. For privacy reasons, these visualizations are computed from traces collected by the authors. The left part of the figure shows the real paths while the right one reflects the estimated paths. The obtained results support that our method is valid as in all three cases the correct entrances and stationary end point was detected by our methods. Given, e.g., the obtained 15 day dataset, our methods help analyze where people enter and end up in the building, and judge if the the paths people currently take through the hospital are optimal, or whether instead means for improved directing of flow would yield improved efficiency or safety.

VIII. CONCLUSIONS

In this paper we have proposed a rich set of features and analysis methods to inform building facility planning

enabling studies of people's behavior in large building complexes utilizing solely measurements of WiFi signals from peoples' devices. To this end, we have addressed the challenges coming with the complexity of the chosen environment. To the best of our knowledge, this is the first study of its type which addresses hospital complexes. The proposed analysis methods include a method to estimate when and where users (respectively their mobile devices) enter and leave buildings. This addresses shortcoming usually inherent in the WiFi-based tracking and offers several possibilities, e.g., to analyze the flow of people from the specific moment they enter a building. We provide a labeling method for differentiating people at the hospital into roles such as staff and visitors, allowing us to obtain information about behavior of respective groups inside the buildings. While we are conscious that this kind of classification is highly dependent on the environment to be analyzed, we also demonstrate that—making use of machine learning tools and of only some domain expertise—we can estimate the role of persons. A limitation of the presented evaluation was that given the privacy concerns of the used test bed we could not collect comprehensive ground truth data for the behavior classification. Nonetheless, we achieved the central goal to provide realistic information that reflects realistically the behavior of, e.g., hospital staff or visitors who make use of the facilities and services offered. Thus, the proposed methods can provide valuable sources of information, e.g. regarding building, path and service utilization, for supporting hospital planning activities.

Building on presented results, for future work we plan to evaluate analysis methods for further aspects of human behavior, consider the development of privacy protecting methods to enable gathering of labeled data in hospital environments, and conduct further evaluations of the visual analysis tools in cooperation with hospital planners.

ACKNOWLEDGMENT

The authors acknowledge the support granted by the *Danish Advanced Technology Foundation* under J.nr. 076-2011-3. Antonio J. Ruiz-Ruiz was partly supported by the Seneca Foundation under the Seneca Program 2009, the Spanish MINECO, as well as EC FEDER funds, under grant TIN2012-38341-C04-03".

REFERENCES

[1] N. Edwards and A. Harrison, "The hospital of the future: planning hospitals with limited evidence. a research and policy problem," *British Medical Journal, 319: 1361*, 1999.

[2] R. B. Bachouch, A. Guinet, and S. Hajri-Gabouj, "An integer linear model for hospital bed planning," *International Journal of Production Economics*, vol. 140, no. 2, pp. 833 – 843, 2012.

[3] G. Ma and E. Demeulemeester, "A multilevel integrative approach to hospital case mix and capacity planning," *Computers and Operations Research*, vol. 40, no. 9, pp. 2198 – 2207, 2013.

[4] P. VanBerkel and J. Blake, "A comprehensive simulation for wait time reduction and capacity planning applied in general surgery," *Health Care Management Science*, vol. 10, no. 4, pp. 373–385, 2007.

[5] A. Marshall, C. Vasilakis, and E. El-Darzi, "Length of stay-based patient flow models: Recent developments and future directions," *Health Care Management Science*, vol. 8, no. 3, pp. 213–220, 2005.

[6] B. Rechel, S. Wright, J. Barlow, and M. McKee, "Hospital capacity planning: from measuring stocks to modelling flows," *Bulletin of the World Health Organization*, 2010.

[7] A. B. M. Musa and J. Eriksson, "Tracking unmodified smartphones using wi-fi monitors," in *Proc. of SenSys*. ACM, 2012, pp. 281–294.

[8] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network," in *Proc. of MobiCom*. ACM, 2004, pp. 187–201.

[9] M. Balazinska and P. Castro, "Characterizing mobility and network usage in a corporate wireless local-area network," in *Proc. of MobiSys*. ACM, 2003, pp. 303–316.

[10] M. Afanasyev, T. Chen, G. M. Voelker, and A. C. Snoeren, "Usage patterns in an urban wifi network," *IEEE/ACM Trans. Netw.*, vol. 18, no. 5, pp. 1359–1372, 2010.

[11] E. O'Neill, V. Kostakos, and T. e. a. Kindberg, "Instrumenting the city: Developing methods for observing and understanding the digital cityscape," in *Proc. of UbiComp*. ACM, 2006, pp. 315–332.

[12] A. Millonig and G. Gartner, "Identifying motion and interest patterns of shoppers for developing personalised wayfinding tools," *J. Location Based Services*, vol. 5, no. 1, pp. 3–21, 2011.

[13] B. E. Moore, S. Ali, R. Mehran, and M. Shah, "Visual crowd surveillance through a hydrodynamics lens," *Commun. ACM*, vol. 54, no. 12, pp. 64–73, 2011.

[14] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: Segmenting space through digital signatures," *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 78–84, 2010.

[15] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao, "Automatically characterizing places with opportunistic crowdsensing using smartphones," in *Proc. of UbiComp*. ACM, 2012, pp. 481–490.

[16] L. Vu, Q. Do, and K. Nahrstedt, "Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace," in *Proc. of IEEE PerCom*, 2011, pp. 54–62.

[17] M. B. Kjærgaard, M. Wirz, D. Roggen, and G. Tröster, "Mobile Sensing of Pedestrian Flocks in Indoor Environments using WiFi Signals," in *Proc. of IEEE PerCom*, 2012.

[18] M. B. Kjærgaard, M. Wirz, D. Roggen, and G. Tröster, "Detecting pedestrian flocks by fusion of multi-modal sensors in mobile phones," in *Proc. of UbiComp*. ACM, 2012, pp. 240–249.

[19] M. B. Kjærgaard, H. Blunck, and M. e. a. Wüstenberg, "Time-lag method for detecting following and leadership behavior of pedestrians from mobile sensing data," in *Proc. of IEEE PerCom*, 2013.

[20] J. K. Laurila, D. Gatica-Perez, I. Aad, and J. e. a. Blom, "From big smartphone data to worldwide research: The mobile data challenge," *Pervasive and Mobile Computing*, 2013.

[21] M. B. Kjærgaard, M. V. Krarup, A. Stisen, T. S. Prentow, H. Blunck, K. Grønbæk, and C. S. Jensen, "Indoor positioning using wi-fi–how well is the problem understood?" in *Proc. of IPIN*, 2013.

[22] A. LaMarca, Y. Chawathe, S. Consolvo, and J. H. et al., "Place lab: Device positioning using radio beacons in the wild," in *Proc of IEEE PerCom*, 2005, pp. 116–133.

[23] P. Bahl and V. N. Padmanabhan, "Radar: an in-building rf-based user location and tracking system," in *Proc of IEEE InfoCom*, 2000, pp. 775–784.

[24] H. Blunck, M. B. Kjærgaard, and T. S. Toftegaard, "Sensing and classifying impairments of gps reception on mobile devices," in *Pervasive*. Springer, 2011, pp. 350–367.

[25] P. Zhou, Y. Zheng, Z. Li, M. Li, and G. Shen, "Iodetector: a generic service for indoor outdoor detection," in *SenSys*. ACM, 2012, pp. 113–126.

[26] M. B. Kjærgaard and C. V. Munk, "Hyperbolic location fingerprinting: A calibration-free solution for handling differences in signal strength," in *Proc. of IEEE PerCom*, 2008, pp. 110–116.

[27] M. Hall, E. Frank, G. Holmes, and et al., "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[28] T. King and M. B. Kjærgaard, "Composcan: adaptive scanning for efficient concurrent communications and positioning with 802.11," in *Proc. of MobiSys*. ACM, 2008, pp. 67–80.

[29] I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub, 2005.