Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 110 (2017) 182–189

Procedia Computer Science

www.elsevier.com/locate/procedia

The 14th International Conference on Mobile Systems and Pervasive Computing (MobiSPC-2017)

Human behavior monitoring using a passive indoor positioning system: a case study in a SME

Pedro E. Lopez-de-Teruel[a], Felix J. Garcia[a,*], Oscar Canovas[a], Ruben Gonzalez[b], Jose A. Carrasco[b]

[a]*University of Murcia, Facultad de Informatica, Murcia 30100, Spain*
[b]*Diwow, Pol. ElDulze, Murcia 30739, Spain*

Abstract

The widespread use of mobile devices such as laptops, smartphones or tablets enables new opportunities and services in the field of pervasive computing and sensing. In particular, monitoring the activity of those devices in indoor working environments enables new methods to address some issues related to energy consumption or employees' wellness. However, it is possible also to infer data about the behavioral pattern of the staff in order to increase productivity, for example identifying anomalies in working teams or unusual behaviors of some employees. In this paper we present a case study for a SME (Small Medium Enterprise) with 20 employees distributed in 5 working teams that develop their daily work in a two-floors building with a WiFi-based passive localization system. An initial analysis of the 802.11 radio signals collected by the system determines, with a high accuracy rate, which mobile devices among the thousands of recorded MAC addresses belong to employees. Additionally, making use of the localization engine, we are able to identify working patterns for the different working teams that, consequently, open the way for implementing efficient anomaly detection techniques.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Indoor localization; behavior analysis; machine learning; case study

1. Introduction and motivation

The proliferation of mobile devices, in particular in micro-enterprises as well as small and medium-sized enterprises (SMEs), is enabling new possibilities in order to infer information about the behavior and activities of the users carrying those devices[1]. In this sense, a relevant research line is related to monitor people behavior in order to reduce energy consumption or improve wellness of workers[2]. Moreover, there is a need to understand working patterns, which can be helpful for improving productivity in an environment with changing work patterns.

Indoor positioning systems (IPS) provide specific locations of mobile devices inside a building traditionally using their radio signals. In particular, some passive IPS are implemented deploying 802.11 monitoring equipment in the

* Corresponding author. Tel.: +34-868-88-8513 ; fax: +34-868-88-4311.
E-mail address: fgarcia@um.es

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.06.076

areas of interest to perform passive localization, that is, without requiring the explicit collaboration of the users. The passive localization considers that mobile devices periodically scan 802.11 channels for access points, which typically involves the transmission of probe messages at different broadcast rate depending on the device. Additionally, those devices send data frames if they are connected to some existing wireless network. Both cases allow the collection of the generated radio signals by special monitors. However, this does not necessarily imply the deployment of new elements, since we can make use of existing hardware in order to add the monitoring functionality. The IPS deployed for the scenario presented in this paper is similar to the one described in a previous publication in[3].

With the maturity of pervasive computing and sensing techniques, extensive research is being conducted for understanding human behavior using different sensing techniques[4]. Several research works have been carried out for indoor localization in order to offer diverse solutions, e.g. monitoring elderly behavior[5], anomaly detection in smart home[6], or occupancy information in universities[3,7]. However, to the best of our knowledge, there are few research works focused on analyzing datasets collected from IPSs running in SMEs, just as in[8] is performed for other different public environments. Among the advantages of monitoring employees, it would be possible to identify deviations in the daily routines, to manage the productivity making use of detailed information and to find potential violations of company policies. In this paper we present an experimental passive localization system which has been deployed in an office building for 1 month. This scenario has been defined mainly for coarse-grained classification purposes, that is, to obtain a room-level location of the mobile devices which belong to the employees.

During the testing period, we have detected more than 45,000 different MAC addresses and more than 1 million signals were collected each day from the monitors, since the system has been able to collect frames from devices in the proximity of the building, which is close to a road. Consequently, the dataset comes from heterogeneous devices generating signals with different RSSI and temporal patterns. This paper proposes to use machine learning techniques to analyze the collected data. Our case study shows how these techniques can be applied in finding working patterns. One of the first contributions of this work is to provide a method to differentiate which of those thousands of MAC addresses correspond to employees and which ones must be discarded. As we will see, as part of our analysis, we will obtain several data features for each collected device and we will determine which features are more suitable to accomplish this first selection task. The second contribution will be the use of techniques for the semi-supervised generation of clusters that will be related to the different working teams. It will be necessary to make use of advanced features which depend on the information obtained from the IPS, such as frequent zones where the employees tend to stay. Those clusters can be then analyzed in order to detect whether particular devices related to particular employees tend to exhibit an anomalous behavior in relation to the corresponding working team.

The rest of the paper is structured as follows. Section 2 presents a characterization of the studied scenario. Section 3 provides an overview of the passive localization system. Section 4 includes a description of the main features extracted from the observations. Section 5 presents the experimental analysis. Finally, conclusions and future work are drawn in Section 6.

2. Studied scenario

For our case study, we deployed our IPS in the head office building of a SME, called AngelPlus[1], that is a custom software development company. The building is distributed in 2 floors of 300 squared meters each and composed of several rooms that are dedicated to offices, technical services, archive, data center and meeting for discussion or commercial purposes. According to areas of interest, we have defined 4 areas (separated by red dotted lines in Figure 1) which refer to the different working zones. The zones B1 and B2 are located on the first floor, and the zones A1 and A2 on the second floor. Both floors are connected by one staircase close to the front door in the zone B2, and another staircase in the zone B1, allowing employees a great mobility. In order to collect radio signals, we have installed 7 wireless monitors (called AP1-AP7 in Figure 1).

In relation to working teams, AngelPlus has a typical organization that consists of five departments or functions: software deployment, technical support, commercial, administration, and management. The software working group, also called SAS, is composed by 14 developers usually located in the zone A1. The technical support group, also called SAT, is provided by 2 technicians in the zone B1. The commercial department with 5 people and administration

[1] http://www.angelplus.es

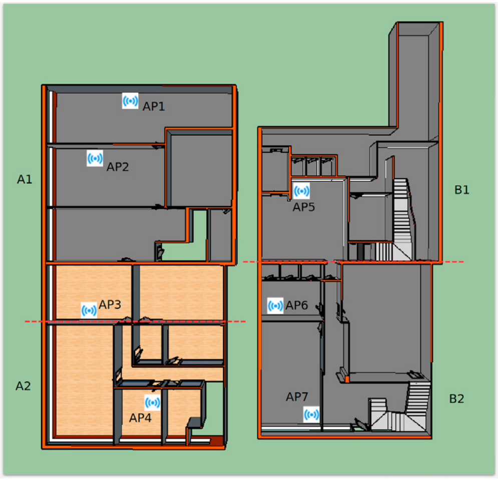

Fig. 1: Floor plans of AngelPlus head office building. (Left) Second floor with zones A1 and A2. (Right) First floor with zones B1 and B2.

department with other 2 workers share the zone B2. Finally, the zone A2 is used by the company manager and also has several meeting rooms. All the 24 employees have the same work schedule; starting work around 9:00 and leaving a bit after 18:00 with about an hour lunch at 14:00, although commercial staff and manager usually visit clients or potential clients and therefore their work schedule is very variable.

Other important aspect of the scenario is the localization of the building. It is located in an industrial park where a large number of people walk or drive by. Moreover, the building is attached to other buildings where other people could work and carry mobile devices.

3. Passive indoor localization system

The radio signals and localizations of mobile devices are captured by a passive IPS implemented by authors and tested in other scenarios [3,9]. With these tests and experiments, we showed the suitability of our proposal for occupancy-based applications. Our IPS is based on wireless localization schemes, where the Received Signal Strength (RSS) is the main source of data for location determination [10]. In these methods, the localization process is usually divided into two phases, namely, the *training phase* and the *online operation*, each presenting their own implementation issues. The training phase involves a site survey process in which the RSSIs at every point of interest is recorded in order to build the fingerprinting database. As for the online operation phase, considering a passive approach, monitoring equipment use the fingerprinting database to infer the position and track unmodified mobile devices in the areas of interest.

In a broad sense, a monitor can be any hardware element running software able to capture 802.11 traffic, although in AngelPlus we have used dedicated access points (AP) with adapted firmware. Our monitors rely only on monitoring the frames normally transmitted by user devices as part of their usual 802.11 connections or active scanning periods. In order to perform its capturing process, each of our monitors simply scans periodically the different 802.11 channels following a plain round-robin schedule. Parameters of this continuous process, such as the scan time for each channel, the set of channels to scan, or the maximum amount of time before a monitor transmits the collected information to the server, are fully configurable. For each captured packet the only information which is used is the MAC address of the emitting mobile device –which is key-hashed for privacy reasons–, the Recieved Signal Strength Indicator (RSSI) and the corresponding timestamp.

For training phase, our monitors can also be configured to act as conventional AP in order to create a fingerprint of radio signals for every zone of interest. Using a training mobile device, we obtain the RSSI values of the beacon frames transmitted by our monitors in AP mode and those observations are tagged (x, y) using a local coordinate system. Our operators follow the indications of a training application that provides accurate geoposition of the walking paths to cover and the required scanning time. For our case study, the training database is composed by 3,129 raw RSSI captured signals, which required about 30 minutes of training time to cover the whole building.

As we have previously mentioned, our work is focused on passive systems. In contrast with active systems, which represented the most frequent approach in the past for designing indoor positioning systems [11], we can summarize the following set of intrinsic features:

- There is no need for special software installed on the mobile devices to be tracked.
- They require the deployment of special purpose devices usually called monitors.
- Estimations and other calculations are usually performed by an external server.
- Mobile devices are assumed to be heterogeneous and an uncalibrated approach is required.
- Traffic patterns are unpredictable since there is no (or minimal) control over the mobile devices.

Regarding the mobile devices being monitored, our sensing system has to provide a flexible characterization for them all. We assume that they will show a wide variety of hardware, WiFi interfaces, antennas, operating systems, and the like. Consequently, they will produce signals with very different strength and temporal patterns. As we want to use the frames normally transmitted by the devices as carried by any type of users during their daily routines, we cannot impose any restrictions on the specific device that each user must have, or make any assumptions on its current state (whether it is switched on or in a low power suspended state, if it is connected or not to any WiFi access point, etc.). In particular, Angelplus mobile devices provided for our experimental analysis are a total of 24 devices among smartphones, tablets and laptops of various manufacturers and suppliers. Note that the real number of devices is larger because several employees have not provided information about some of their additional personal mobile devices.

Finally, a central server will be in charge of hosting the localization and tracker engine itself. A central element of this engine will be a database containing both the fingerprinting data collected during the training phase, as well as the continuously updated information of the captures sent to it by the monitors. On the other hand, the server will be also in charge of running the localization software responsible for calculating occupancy and positioning information when required. In order to do that, the server provides an API that will be used by higher-level location-based services.

The reader is referred to [3] and [12] for further details on our passive localization engine implementation.

4. Behavior features

Multiple features can be extracted from the collected radio observations. If we analyse the data captured by monitoring equipment, there are four useful items: timestamp, anonymized MAC address of mobile device, monitor identifier, and RSSI value. Monitor identifiers, ranging from AP1 to AP7, are the identities of the seven wireless capturing cards. An RSSI is a measure of the power level that the monitor is receiving from the mobile device. RSSI is the relative signal strength expressed as dBm between 0 and -100. As a general example, a good signal would be -50, a reasonable would be -75, and a bad one would be -90, while -100 would provide no service at all.

Considering these items, we define a set of *raw behaviour features*, i.e. only using the data captured by the monitors without using any positioning algorithm. This set of raw behaviour features for each mobile device are:

- Temporal features: number of readings for each weekday (from Monday to Sunday) and for each part of the day (morning, afternoon and night).
- Accounting features: global number of readings for each device and number of days a device was registered.
- Visibility features: number of readings for each monitor, and maximum number of monitors that simultaneously registered the mobile device.

Additionally, we can calculate more features using a positioning engine that provides location data about the mobile devices which are located inside the building. For example, a session consists of a timestamped list of visited zones. The start time of the first zone is the session start time and the end time of the last zone is the session end time. The session time between zones is seamless, and two different sessions cannot be overlapped for the same mobile device.

Using this calculated dataset, we define other set of features, namely, *session behaviour features* that take particular advantage of the positioning information generated by IPS. This set of session features for each mobile device are:

- Temporal features: number of sessions for each weekday (from Monday to Sunday) and for each part of the day (morning, afternoon and night).

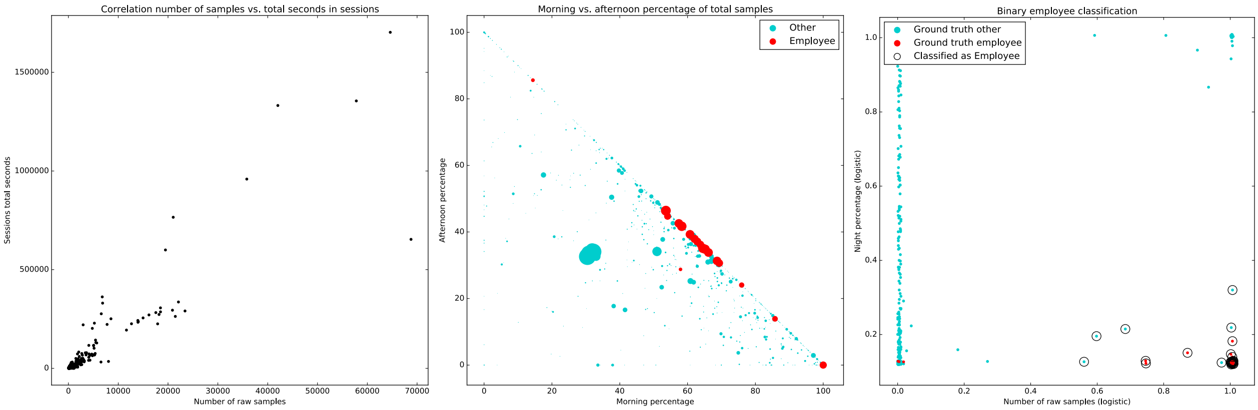

Fig. 2: (Left) Correlation between real time in seconds (estimated after processing by sessions) and original (unprocessed) raw samples. (Middle) Percentage of morning vs. afternoon raw (unprocessed) samples. (Right) Binary classification employee MAC / non employee MAC.

- Spatial features: total time (in seconds) in each zone, maximum number of changes between zones in a session, and mobility factor (changes of zone in relation to total session time).
- Accounting features: sum of days when the mobile device was tracked, and total tracking time (in seconds).

These behaviour features are the basis of the experimental analysis process based on machine learning techniques that we present in the next section.

5. Experimental analysis

We have performed a set of experiments to validate both the first stage of binary classification employee/not employee (using the *raw features*) and the second stage multiclass classification by type of employee (using the *session features*). In both cases we compute the values of *precision*, *recall* and *f1-score* [13] of the obtained classifier, as well as the *homogeneity*, *completeness* and *V-measure* [14] as a measure of the goodness of the clustering, in this last case just for the multilabel classification. In all cases the corresponding figures of merit take values in the interval [0.0; 1.0], with 1.0 corresponding to perfect classification/clustering.

5.1. Employee / non employee classification based on raw behaviour features

In order to implement our first stage classifier, we must select a good subset of discriminant raw features. In Figure 2 (left) we show that the total number of raw (unprocessed) samples for a given MAC in the studied period exhibits a high correlation with the real time in seconds which was estimated after processing those samples to get the corresponding sessions. This allows us to use that simple raw feature as a good approximation of acummulated time inside the target scenario –a very useful feature to determine if the corresponding MAC belongs to an employee–, without the need to further process the raw information obtained by the passive monitoring system. On the other hand, in Figure 2 (middle), where the size of each sample is made proportional to the aforementioned number of raw samples for each MAC, we can also observe that, except for some isolated employee, almost all of them (red points in the figure) fall in the diagonal line *%morning*+*%afternoon*=100, i.e., none of them is detected in the building by night. Though it is true that there are also many other non employee MACs along this line (cyan points in the same diagonal), they are clearly smaller in size (meaning that they do not spend as much time inside the building as employees). In contrast, there are also large cyan points outside the aforementioned diagonal (in fact, some of the largest cyan circles are in that situation). This accounts for the fact that they correspond to MACs of non mobile devices which generate WiFi traffic and that tend to remain switched on all day long (we call these 24/7 MACs). This is also an important fact that we had to take into account in our binary classifier, in order not to confuse these 24/7 MACs with employees.

In order to convert both of these numerical features (total number of raw samples, and night time percentage) to normalized values in the $[0, 1]$ interval, corresponding to certainty values with respect to the associated boolean predicates *"Frequently detected MAC"* and *"Frequently nightly detected MAC"*, we used a standard binary logistic regression procedure [15]. The applied logistic function to each feature is $\sigma(x; c, s) = \frac{1}{1+e^{-\frac{x-c}{s}}}$, where c is the midpoint of the logistic regression (maximum uncertainty with respect to the boolean output value of the feature) and s is a scale factor which determines the speed at which greater/smaller values asymptotically approach the certainty values true/false, respectively. We chose reasonable values for s and c in both features by looking at both left and middle plots of Fig. 2, considering that *a)* more than 5000 raw samples, corresponding to approximately 24 hours = 3 days of work is a good frontier to consider a MAC as "frequently detected inside the building" and *b)* a night percentage appreciably different from zero (say, for example., > 5% of total time) corresponded to a "nightly detected MAC". The obtained binary classification results (using a simple nearest neighbour classifier based on the two mentioned logistic features, which was trained in a very light semisupervised fashion using just three known MACs corresponding to a known employee, a typical passing by car, and a known 24/7 type MAC) are summarized in both Fig. 2 (right) and Table 1. We can see that the classifier behaves almost perfectly ($\sim$ 0.999) in the (easier, and much more numerous) case of non-employees, while still obtaining rather acceptable accuracy values of $\sim$ 0.8 for the case of employees. We must remember also here that, as was stated in section 3, not all personal mobile devices MACs of all employees were revealed as ground truth for this experimental validation procedure, so they could certainly account for a good amount of these remaining 20% errors.

Table 1: Employee / non employee binary classification results.

	Precision	*Recall*	*F1-score*	*Support*
Non employee	0.999	0.999	0.999	49000
Employee	0.741	0.833	0.784	24
Average / total	0.999	0.999	0.999	49024

5.2. Classification by type of employee based on session behaviour features

Regarding feature selection and preprocessing for this second stage classifier, we use standard PCA dimensionality [16] reduction to 2D pairs of features to help in visualization. Fig. 3 (left) shows the first two PCA components of the analysis of the seven features corresponding to the total session time spent by each MAC for the seven days of a week. We start to perceive a pattern of aggregation by colors, though with clear areas of overlap between administratives, SAS and commercials. Still, it is interesting to note again large gray circles (corresponding to inside 24/7 MACs, and thus corresponding to non-employee devices) far from the main variation mode in the 2D line along which employees can be found. Finally, both a SAT and the manager are the only employees that tend to go to work also on saturday mornings, while the rest of employees only go from monday to friday. This translates in them to be also out of the dominant variation line, though not as much as the 24/7 devices in gray at the bottom of the figure.

The other two graphics in Fig. 3 show again a PCA 2D analysis, this time of the absolute session time (middle) and session time percentage (right) spent by each MAC in the given five building zones. This time, same type of employees tend to appear clearly grouped, perhaps more tightly in the percentage case (thus absorbing the absolute time variability which translates in more elongated clusters in the middle figure). Moreover, admins and commercials tend to appear together (as their usual working places are next to each other, both in zone B2), and something similar occurs with SASs and SATs (which tend to stay between zones A1 and B1). The manager appears in a well separated cluster, corresponding to a clearly different zone (A3). We also found that the *mobility* related features were not too useful to aid in class separation by type of employee. Still, certain types of employees, such as the administrative staff, tended to present a much smaller mobility than commercials, SASs and SATs, though we do not present those additional results here due to lack of space.

Based on these features, and using again a semisupervised approach, we implemented a simple nearest neighbour based clustering technique which used one sample prototype per each type of employee. Of course, the quality of the obtained classification results depended on the exact set of chosen prototypes (see for example Fig. 4 for two different examples, one with a worst performing initial labelling, the other with a much better one). The obtained results for precision, recall, F1-score, homogeneity, completeness and V-measure for the first (worst) case was of, respectively,

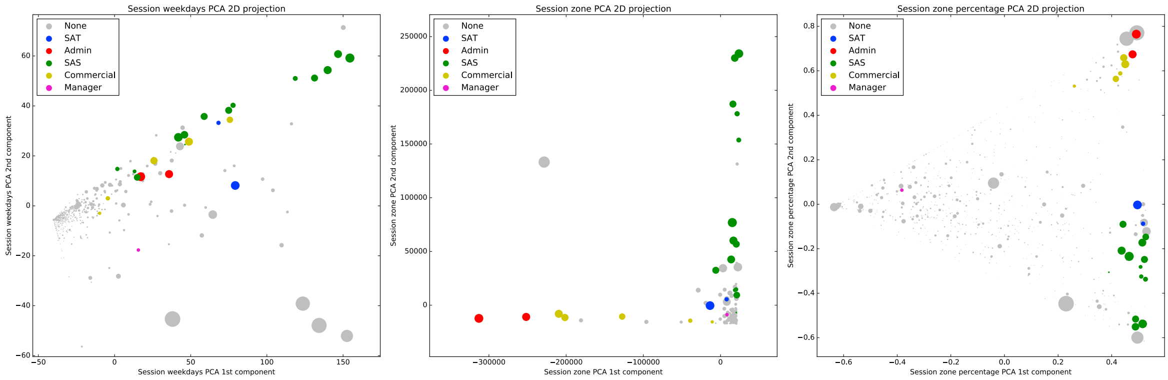

Fig. 3: Different subsets of features: (Left) First two PCA components of total session time by week days. (Middle) First two PCA components of total session time by building zones. (Right) Same as in middle plot, but using time percentage by zone instead of total session time.

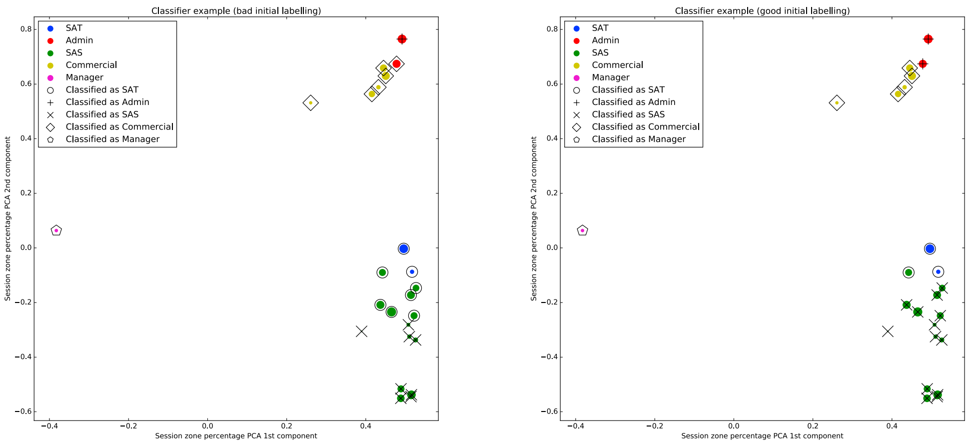

Fig. 4: Two different classifiers obtained when using a bad (left) and good (right) initial set of samples to perform the semisupervised clustering.

0.90, 0.71, 0.74, 0.75, 0.66 and 0.70, while the corresponding values for the second (best case) was of 0.97, 0.96, 0.96, 0.93, 0.88, and 0.91.

To cope with this variability, and in order to obtain a more informative evaluation of the procedure, independent of the set of chosen prototypes, we evaluated all the possible semisupervised clusterings using all possible $2 \times 2 \times 14 \times 5 \times 1$ (2 SAT, 2 administratives, 14 SAS, 5 commercials, and 1 manager), for a total of 280 combinations of unique 5 prototypes (SAT, administratives, SAS, commercials, and manager). Tables 2 and 3, finally, show the corresponding multilabel classification results, as well as the corresponding accumulated and normalized confusion matrix, using all these possible initial minimal sets of prototypes as seeds to perform the nearest neighbour semisupervised clustering. In all cases, both mean and standard deviation for the respective values of the 280 tests performed are shown.

Table 2: Detailed classification by type of employee results (using all possible initial minimal sets of prototypes; see text). Associated *homogeneity*, *completeness* and *V-measure* scores were 0.82 ± 0.07, 0.74 ± 0.08 and 0.78 ± 0.08, respectively.

	Precision	*Recall*	*F1-score*	*Support*
SAT	0.43 ± 0.20	0.96 ± 0.13	0.57 ± 0.18	2
Admin	0.79 ± 0.27	0.90 ± 0.20	0.79 ± 0.19	2
SAS	0.99 ± 0.02	0.74 ± 0.19	0.83 ± 0.14	14
Commercial	0.97 ± 0.07	0.80 ± 0.28	0.84 ± 0.21	5
Manager	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1
Average / total	0.93±0.19	0.80±0.22	0.82±0.18	24

Table 3: Confusion matrix for classification by type of employee (using all possible initial minimal sets of prototypes; see text).

	SAT	Admin	SAS	Commercial	Manager	Total
SAT	**96%** (540)	0%	4% (20)	0%	0%	280*2
Admin	0%	**0.90%** (504)	0%	10% (56)	0%	280*2
SAS	26% (1010)	0%	**74%** (2910)	0%	0%	280*14
Commercial	0%	20% (280)	0%	**80%** (1120)	0%	280*5
Manager	0%	0%	0%	0%	**100%** (280)	280*1

6. Conclusions and future directions

In this paper we have presented a case study for monitoring human behavior in SMEs based on a passive indoor localization system, by focusing on behavioral patterns that are constructed according to the occupancy of the staff in specific areas of interest. We have studied several type of features, based on both raw samples (as directly obtained by the passive monitoring) and processed features (as postprocessed by a fingerprinting location engine), discussing and analyzing their performance on two respective real-wold cases of application, namely a first stage binary employee/non-employee classification and a second stage multiclass clustering by type of employee. In both cases we have found that accuracies of around 80%-90% can be obtained using a simple nearest neighbour approach.

Future work will be oriented at extending our experimentation to other scenarios, such as shopping malls, university buildings, and other different type of SMEs, where we plan to test other different sets of features to evaluate the possibly very different behaviors of people in these other types of environments.

Acknowledgments

This work was supported by the Spanish MINECO, as well as European Commission FEDER funds, under grant TIN2015-66972-C5-3-R.

References

1. Lockhart, J.W., Pulickal, T., Weiss, G.M.. Applications of mobile activity recognition. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM; 2012, p. 1054–1058.
2. Zenonos, A., Khan, A., Kalogridis, G., Vatsikas, S., Lewis, T., Sooriyabandara, M.. Healthyoffice: Mood recognition at work using smartphones and wearable sensors. In: *PerCom Workshops*. IEEE; 2016, p. 1–6.
3. Lopez-de Teruel, P.E., Garcia, F.J., Canovas, O.. Validating passive localization methods for occupancy sensing systems in wireless environments: a case study. *MobiSPC 2016 - Procedia Computer Science* 2016;**94**:57–64.
4. Atallah, L., Yang, G.Z.. The use of pervasive sensing for behaviour profiling: a survey. *Pervasive and Mobile Computing* 2009;**5**(5):447–464.
5. Barsocchi, P., Cimino, M.G., Ferro, E., Lazzeri, A., Palumbo, F., Vaglini, G.. Monitoring elderly behavior via indoor position-based stigmergy. *Pervasive and Mobile Computing* 2015;**23**:26–42.
6. Bakar, U., Ghayvat, H., Hasanm, S., Mukhopadhyay, S.. Activity and anomaly detection in smart home: A survey. In: *Next Generation Sensors and Systems*. Springer; 2016, p. 191–220.
7. Li, K., Chau, Y., Kanhere, S.. Senseflow: An experimental study of people tracking. In: *ACM Workshop on Real World Wireless Sensor Networks (REALWSN)*. ACM; 2015, p. 31–34.
8. Ruiz-Ruiz, A.J., Blunck, H., Prentow, T.S., Stisen, A., Kjaergaard, M.B.. Analysis methods for extracting knowledge from large-scale wifi monitoring to inform building facility planning. In: *Pervasive Computing and Communications (PerCom)*. IEEE; 2014, p. 130–138.
9. Lopez-de Teruel, P.E., Canovas, O., Garcia, F.J.. Visualization of clusters for indoor positioning based on t-SNE. In: *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. 2016, p. 199–202.
10. He, S., Chan, S.H.G.. Wi-Fi fingerprint-based indoor positioning: Recent advances and comparisons. *IEEE Communications Surveys & Tutorials* 2016;**18**(1):466–490.
11. Koyuncu, H., Yang, S.. A survey of indoor positioning and object locating systems. *International Journal of Computer Science and Network Security* 2010;**10**(5):121–128.
12. Lopez-de Teruel, P., Canovas, O., Garcia, F.J.. Using dimensionality reduction techniques for refining passive indoor positioning systems based on radio fingerprinting. *Sensors* 2017;**17**(4).
13. Powers, D.M.. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies* 2011;**2**(1):37–63.
14. Roseberg, A., Hirschberg, J.. V-measure: A conditional entropy-based external cluster evaluation measure. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007, p. 410–420.
15. Hilbe, J.M.. *Logistic Regression Models*. CRC Press; 2009.
16. Jolliffe, I.. *Principal Component Analysis*. Springer; 2^{nd} ed.; 2002.